UniSA: Unified Generative Framework for Sentiment Analysis

Zaijing Li*
Central South University
lzj14011@gmail.com

Ting-En Lin
Alibaba Group
ting-en.lte@alibaba-inc.com

Yuchuan Wu
Alibaba Group
shengxiu.wyc@alibaba-inc.com

Meng Liu
Shandong Jianzhu University
mengliu.sdu@gmail.com

Fengxiao Tang[†]
Central South University
tangfengxiao@csu.edu.cn

Ming Zhao[†]
Central South University
meanzhao@csu.edu.cn

Yongbin Li[†]
Alibaba Group
shuide.lyb@alibaba-inc.com

ABSTRACT

Sentiment analysis is a crucial task that aims to understand people's emotional states and predict emotional categories based on multimodal information. It consists of several subtasks, such as emotion recognition in conversation (ERC), aspect-based sentiment analysis (ABSA), and multimodal sentiment analysis (MSA). However, unifying all subtasks in sentiment analysis presents numerous challenges, including modality alignment, unified input/output forms, and dataset bias. To address these challenges, we propose a Task-Specific Prompt method to jointly model subtasks and introduce a multimodal generative framework called UniSA. Additionally, we organize the benchmark datasets of main subtasks into a new Sentiment Analysis Evaluation benchmark, SAEval. We design novel pre-training tasks and training methods to enable the model to learn generic sentiment knowledge among subtasks to improve the model's multimodal sentiment perception ability. Our experimental results show that UniSA performs comparably to the state-of-the-art on all subtasks and generalizes well to various subtasks in sentiment analysis.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; **Multimedia information systems**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

sentiment analysis, multimodal information, unified framework

ACM Reference Format:
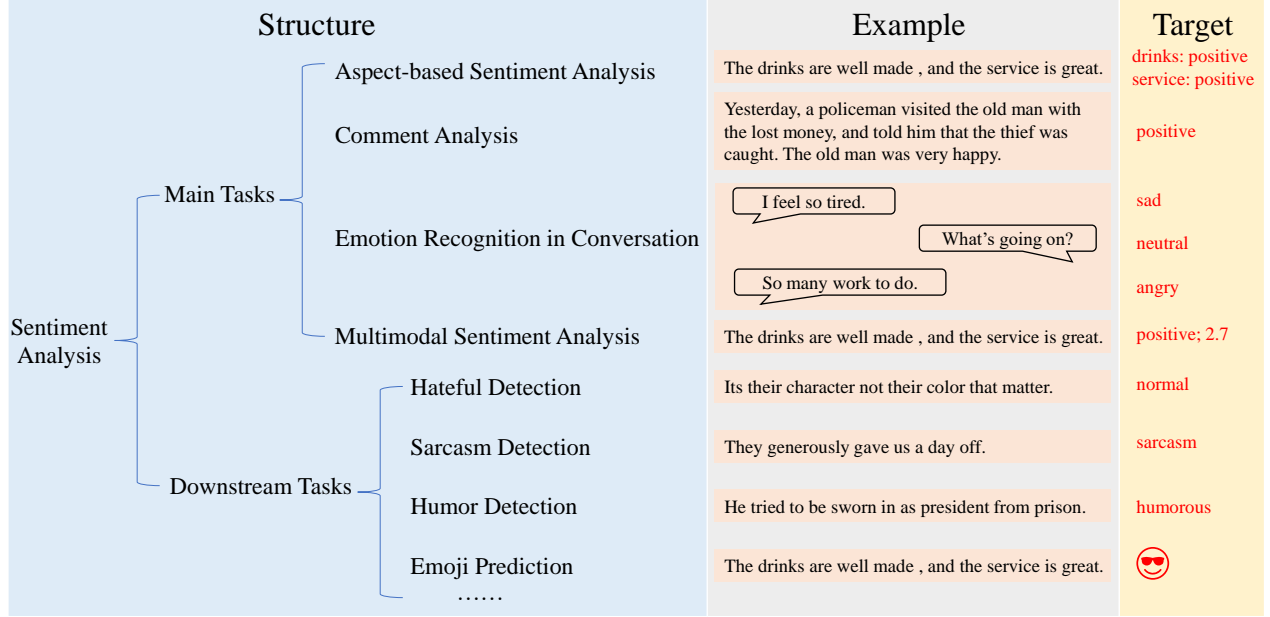Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. UniSA: Unified Generative Framework for Sentiment Analysis. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612336

---

1 INTRODUCTION

Sentiment analysis is a discipline that leverages multimodal data to extract human opinions and comments, as well as comprehend and categorize human emotions. Generalized sentiment analysis encompasses a plethora of subtasks, such as emotion recognition in conversation (ERC), aspect-based sentiment analysis (ABSA), and multimodal sentiment analysis (MSA). Initially, research focused solely on individual subtasks; nevertheless, it has become evident that there is an interrelated sentiment knowledge among these subtasks. Hence, integrating all subtasks into a single model to enhance the sentiment understanding ability of the model has emerged as a significant objective. Following the lead of unified multi-task modeling in other domains, recent studies have explored the potential of jointly modeling some subtasks, e.g., Hu et al. [18] jointly modeled the ERC and MSA to boost the performance of both tasks, Yan et al. [70] converted all ABSA subtasks a unified generative formulation, which yielding encouraging results. Nevertheless, no research has yet been conducted on the joint modeling of all sentiment analysis subtasks (ERC, MSA, ABSA, etc.) as a single research object.

The unified modeling of all subtasks of sentiment analysis presents three primary challenges: 1) **Format**. The input formats and analysis views of each subtask vary. For example, MSA analyzes emotional tendencies based on single-turn utterances, ERC comprehensively assesses speaker emotions through contextual information in dialogues, and ABSA extracts attribute words from utterances and judges emotional tendencies based on those words. Jointly training these subtasks, each with different input and output formats, is the first challenge. 2) **Alignment**. Some subtasks use multimodal data (e.g., ERC), while others use single-modal data (e.g., speech emotion recognition). The data formats and representations for different modalities (text, acoustic, visual, etc.) differ, and each modality expresses emotional information in its unique way. For example, text modality uses sentiment words, attribute words, modal words, and negative words to express sentiments, while the acoustic modality mainly uses acoustic parameters such as intensity, pitch, speaking rate, and pauses to express the speaker's emotional fluctuations.

Structure	Example	Target
Aspect-based Sentiment Analysis	The drinks are well made , and the service is great.	drinks: positive service: positive
Comment Analysis	Yesterday, a policeman visited the old man with the lost money, and told him that the thief was caught. The old man was very happy.	positive
Emotion Recognition in Conversation	I feel so tired. / What's going on? / So many work to do.	sad neutral angry
Multimodal Sentiment Analysis	The drinks are well made , and the service is great.	positive; 2.7
Hateful Detection	Its their character not their color that matter.	normal
Sarcasm Detection	They generously gave us a day off.	sarcasm
Humor Detection	He tried to be sworn in as president from prison.	humorous
Emoji Prediction	The drinks are well made , and the service is great.	😎

Figure 1: Subtasks of sentiment analysis are categorized into main tasks and downstream tasks based on their relevance to the emotion of humans.

The visual modality expresses human emotion through facial expressions, body posture, and eye gaze. Aligning the emotional information across modalities presents the second challenge. 3) **Bias.** Sentiment analysis is a highly subjective task, and ensuring that the model learns universal human sentiment knowledge while being less affected by subjectivity bias is the third challenge. Additionally, dataset annotation bias may affect the quality of multimodal data with high-quality annotations, making it difficult to train models that can generalize across different datasets.

In response to the challenges of sentiment analysis, we reorganize the sentiment analysis subtasks into two categories, namely main tasks and downstream tasks, based on their relevance to sentiment. As shown in Figure 1, the main tasks, which are the subtasks most correlated with human emotional representation, include ABSA, MSA, ERC, and Comment Analysis (CA). Downstream tasks include tasks related to sentiment analysis but not necessarily detecting human emotion categories, such as irony detection, humor detection, and emoji prediction. Meanwhile, we propose a novel multimodal sentiment analysis framework, named UniSA [1], which takes the first step toward unified modeling of all sentiment analysis main tasks and generalizes to downstream tasks.

Specifically, to tackle the first challenge of unifying input and output forms across different subtasks, we introduce the task-specific prompt method which treats all subtasks as generation tasks and jointly trains them. The second challenge of aligning emotional information across different modalities is addressed by extending the generative Transformer [64] architecture to process multimodal data and proposing the modal mask training method to learn the inter-modality relationship. The third challenge of learning the

difference between subtasks is tackled by introducing dataset embedding to bridge the annotation bias between different datasets. In order to evaluate the performance of our proposed framework, UniSA, we collate benchmark datasets for each subtask and construct a new sentiment analysis benchmark, SAEval, as illustrated in Table 1. The details of this benchmark are presented in Section 3.

The contributions of this paper can be summarized as follows:

- We advance a novel approach to sentiment analysis, UniSA, which unifies all subtasks under a single generative framework. This represents a significant advancement in the field, as no previous work has taken such a comprehensive approach to sentiment analysis.
- We propose novel sentiment-related pre-training tasks that allow the model to learn generic sentiment knowledge across subtasks. Extensive experimental results demonstrate that UniSA performs comparably to the state-of-the-art on all subtasks.
- We curate a benchmark dataset, SAEval, which comprises benchmark datasets for various sentiment analysis subtasks in a unified format, enabling comprehensive and fair evaluation of sentiment analysis models.

2 RELATED WORK

2.1 Sentiment Analysis

We provide a brief overview of the recent advancements in various subfields of sentiment analysis. Aspect-based sentiment analysis (ABSA) is a task that aims to identify the sentiment polarity associated with aspect terms in a single-turn utterance. Previous works

---

[1]The UniSA is available at : https://github.com/dawn0815/UniSA

Table 1: Statistics of our SAEval benchmark, where T, A, and V represent text, acoustic, and visual, respectively.

Task	Task type	Dataset	Modality	Average Length (text)	Total Size	Train+Val Size	Test Size
ABSA	Classification	SemEval-2014	T	16	4.8k	3819	1028
ABSA	Classification	SemEval-2016	T	15	1.3k	1067	326
MSA	Regression	MOSI	T+A+V	12	2K	1513	686
MSA	Regression	MOSEI	T+A+V	20	20k	18197	4659
ERC	Classification	IEMOCAP	T+A+V	12	7k	5758	1622
ERC	Classification	MELD	T+A+V	8	13k	11100	2610
ERC	Classification	DailyDialog	T	15	100k	95239	7740
ERC	Classification	EmoryNLP	T	10	12k	11278	1328
ERC	Classification	EmoWOZ	T	11	167k	74983	8634
CA	Classification	SST-2	T	10	11k	68221	1821
CA	Classification	IMDB	T	233	50k	25000	25000
CA	Classification	Amazon Review	T	21	37m	37m	-

have employed the attention mechanism integrated with LSTM-based neural network models to model the relationship between aspects and their contextual words [33, 37, 66]. Recent research on ABSA has explored the application of sequence-to-sequence learning and pre-trained language models to achieve promising results [26, 36].

Multimodal sentiment analysis involves identifying the speaker's emotion from a single-turn utterance by considering multiple modalities. Early research in this area primarily focused on geometric manipulation in feature spaces [75, 76]. Recent research on multimodal sentiment analysis has emphasized the importance of modal consistency and difference through multi-task joint learning [74] or modal translation [39], leveraging cross-modality and multi-scale modality representation to implement the modal alignment [35, 63].

Comment analysis involves identifying the user's emotion from one or more sentences in a comment. Recent works [57, 68, 72] in this field have used pre-trained language models, such as BERT [5], RoBERTa [34], and XLNet [72], to fine-tune comment datasets and achieve promising results.

Emotion recognition in conversation aims to identify the speaker's emotion from multiple utterances in a conversation [8, 31, 51, 60]. Early research focused on context modeling using GRU [6] models to extract context information and judge the emotion category of the utterance based on the context information [9, 12, 17, 40, 49]. More recent research has introduced GCN models [11] into conversation scene modeling, where each utterance in the conversation [32, 80] is regarded as a node in the graph, and the relationship between utterances constitutes the edge connecting the nodes [10, 19, 25, 59]. The most recent works in this area employ Transformer architecture and self-attention mechanisms to capture contextual information of utterances and achieve state-of-the-art performance in emotion recognition in conversations [20, 27, 30, 41, 58].

2.2 Multi-task Unified Framework

In recent years, multi-task unified architectures have shown great potential and achieved impressive results across various domains [13–15, 45, 52, 73]. Bao et al. [1] presented a unified vision-language pre-trained model that utilizes a modular Transformer network to jointly learn a dual encoder and a fusion encoder. Li et al. [28]

proposed a unified pre-training architecture that can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Zhang et al. [81] proposed a Unified framework for multimodal summarization. In the Named Entity Recognition domain, some works designed unified architectures for various subtasks [24, 71]. Recently, large multimodal models such as ERNIE Bot [65] and GPT-4 [46] have achieved remarkable results and have attracted attention from researchers in various fields.

In the sentiment analysis field, Yan et al. [70] employed an improved BART [23] architecture to solve all ABSA subtasks in an end-to-end framework. Hu et al. [18] proposed a multimodal sentiment knowledge-sharing framework that unifies MSA and ERC tasks from features, labels, and models. These multi-task unified works in various fields support the feasibility of unified modeling for all sentiment analysis subtasks. However, there is currently no end-to-end architecture that can model all subtasks of sentiment analysis.

3 SAEVAL: THE BENCHMARK

To better evaluate the performance of the model on various sentiment analysis tasks, we formalized the benchmark dataset for the main task and constructed a new benchmark, SAEval [2]. This section provides a description of the datasets that constitute the SAEval benchmark and the evaluation metrics used.

3.1 Datasets

As presented in Table 1, the SAEval benchmark includes several datasets from different subtasks of sentiment analysis:

- SemEval-2014 [48] and SemEval-2016 [47] are subtasks of the Semeval Aspect-based Sentiment Analysis challenge. The goal of these subtasks is to identify the sentiment polarity (positive, negative, neutral, conflict) corresponding to all attribute words contained in each sentence.
- MOSI [77] and MOSEI [78] are two widely used multimodal sentiment analysis datasets. The goal of SAEval for these two

---

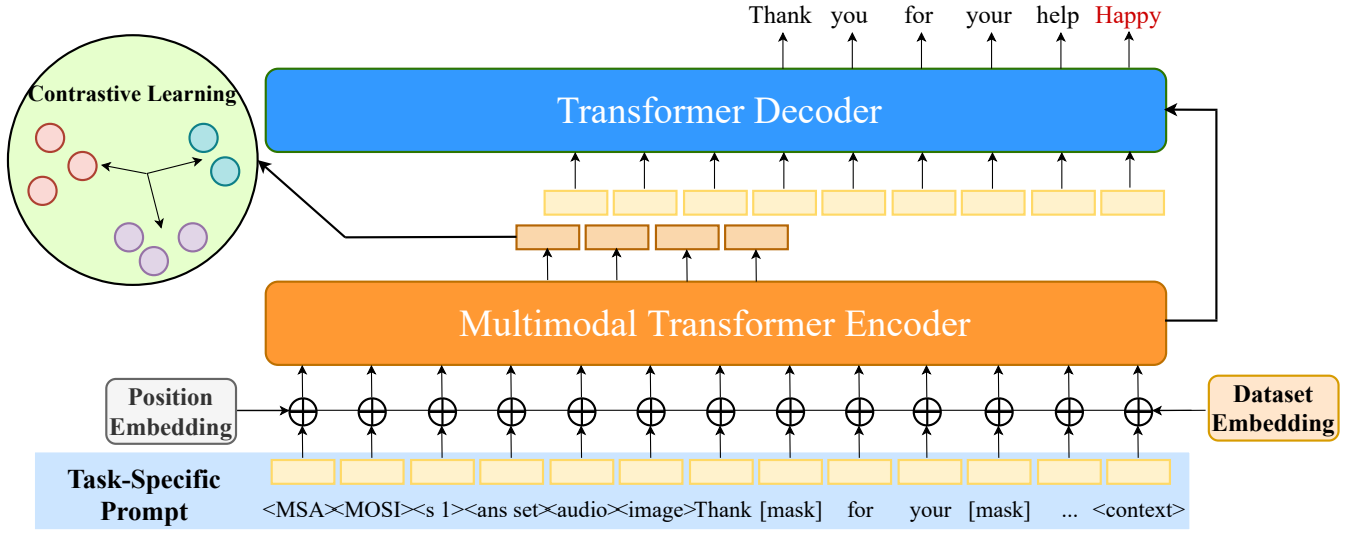[2] The SAEval benchmark : https://github.com/dawn0815/SAEval-Benchmark

Figure 2: Overview of UniSA: a novel approach for unified sentiment analysis modeling.

datasets is to predict the sentiment score, which is a continuous value ranging from -3 to +3, of single-turn utterances by incorporating multiple modalities.

- IEMOCAP [4] and MELD [50] are both datasets for emotion recognition in conversations using multimodal information. The SAEval benchmark uses these datasets to identify the emotion category of each utterance based on the multimodal information and context available.
- EmoryNLP [79], DailyDialog [29], and EmoWOZ [7] are datasets for textual emotion recognition in conversation. The goal of SAEval for these datasets is to identify the emotion category of utterances based on textual information and context.
- SST-2 [61], IMDB [38], and Amazon Review [44] are datasets for comment analysis. The goal of SAEval for these datasets is to identify the sentiment polarity of comments. It is important to note that Amazon Review is only used for the pre-training phase and does not have a test set for evaluation in SAEval.

All datasets are unified and stored in a dictionary format [3]. The dictionary includes keywords, such as "Task Type", "Dataset ID", "Text", "Audio", and "Image". For ERC datasets, additional information such as "Context", "Speaker ID", and "Utterance index" are included to determine the conversation information of the current query.

3.2 Evaluation Metrics

SAEval uses the same evaluation metrics as the original tasks for each subtask. Weighted Accuracy (WA) is used for aspect-based sentiment analysis and comment analysis. Mean Absolute Error (MAE), 7-category Accuracy (ACC-7), and 2-category Accuracy (ACC-2)

are used for multimodal sentiment analysis. WA and weighted F1 score (WF1) are used for emotion recognition in conversation.

As the percentage of neutral categories in the DailyDialog dataset is more than 90%, this dataset is usually measured with neutral categories removed by default, and macro-averaged F1 (MF1) is used as the measure.

4 METHODOLOGY

The architecture of our UniSA is depicted in Figure 2. We adopt a generative Transformer architecture to unify all subtasks of sentiment analysis into generation tasks. Concretely, to handle the cross-modality inputs of visual, acoustic, and text, we modify the original Transformer encoder to a multimodal encoder and introduce a Modal Mask Training method. This method enables the model to learn the relationship between different modalities effectively. We also propose a Task-Specific Prompt method to standardize the input format of all subtasks. Furthermore, to address the bias between datasets, we incorporate a dataset embedding in the input to differentiate between different datasets. This technique helps the model to better understand the characteristics of each dataset and improves its performance on all tasks. We will elaborate on each of them sequentially.

4.1 Problem Formulation

The unified sentiment analysis subtasks aim to process arbitrary data from different modalities, such as text, acoustic, and visual, and output emotion predict results corresponding to a specific subtask. The subtask set includes ABSA, MSA, ERC, and CA, while the modality set includes m_t, m_a, and m_v, corresponding to text, acoustic, and visual modalities, respectively. The task involves both multimodal situations, such as MSA and ERC, as well as unimodal situations, such as ABSA and CA. Moreover, the task must also consider multimodal situations where one or more modalities are missing. Our target is to build a multimodal emotion-aware frame-

---

[3]Please see Appendix A.1 for formatted samples

Table 2: Number of model parameters (in millions) and hyperparameter settings for the fine-tuning phase.

Backbone	Number of Parameters	Learning Rate	Batch Size	Dropout Rate	Epochs	Max Length of Sequence
GPT2-medium	341m	5e-5	32	0.0	15	128
T5-base	221m	5e-5	32	0.0	15	128
BART-base	141m	5e-6	64	0.1	40	600

work based on SAEval, named UniSA, which learns cross-task emotional knowledge through emotion-related pre-training tasks and generalizes to various downstream tasks.

4.2 Task-Specific Prompt

To jointly model different subtasks, we propose Task-Specific Prompt to unify the input streams of all subtasks, and transform all subtasks into generative tasks to unify the output form of subtasks. Task-Specific Prompt comprises three components: task identifier Z, answer set Y, and input streams X. The template L can be represented as:

$$L = \{Z, Y, X\}. \tag{1}$$

As illustrated in Figure 3, the task identifier Z is made up of special tokens, including $< task\_type >, < data\_id >, < speaker\_id >$, etc., which distinguish different subtasks, datasets, and speakers (in conversation). The answer set Y is a specific set of labels for each dataset that guides the model in generating the expected results. The input streams X indicate the text, acoustic, visual, and context of the inputs. Task-Specific Prompt standardizes the input form and guides the model to generate task-specific results according to the answer set.

4.3 Modal Mask Training

As the text modality data is abundant in each subtask of sentiment analysis, while the multimodal data is limited, we propose a Modal Mask Training method to enhance the model's multimodal emotion perception capability. Specifically, given a multimodal input signal $I_i = I_i^t, I_i^a, I_i^v$, where $I_i^m, m \in \{t, a, v\}$ represent unimodal input of time i, we force the masking of one or more of the modal inputs. This transformation yields seven modal settings: $I_i^t, I_i^a, I_i^v, I_i^a + I_i^v, I_i^t + I_i^a, I_i^t + I_i^v, I_i^t + I_i^a + I_i^v$.

However, we found that some modal settings do not work well during training, so we only use the following four modal settings : $I_i^t, I_i^t + I_i^a, I_i^t + I_i^v, I_i^t + I_i^a + I_i^v$. In this way, multimodal data has four different input forms in the training phase, which extends the modal diversity of the training data and can effectively cope with the absence of some modalities in multiple data in real scenarios. This method allows the model to learn the relationship between different modalities effectively and improve its performance on multimodal sentiment analysis tasks where the data is limited.

4.4 Dataset Embedding

To reduce the impact of dataset bias on the model performance, we propose the use of dataset embedding. This technique transforms the dataset indexes into one-hot embeddings, allowing the model to distinguish between different datasets. As sentiment analysis is a subjective task, people's emotional responses to the same sentence

can vary. Additionally, different datasets may be labeled by different annotators, introducing subjective bias between datasets. The dataset embedding helps mitigate this bias, allowing the model to better generalize across different datasets.

4.5 Multimodal Transformer Encoder

The encoder of our model is based on a multi-layer bidirectional Transformer, which is similar to the architecture used by Xing et al. [69]. We adopt a single-stream architecture to jointly train the inputs of different modalities. This architecture allows the model to effectively capture the interactions and dependencies between different modalities and improve its overall performance on all tasks. For visual modality, we use pre-trained MobileNet [16] extracted features as visual embedding, for acoustic modality, we use librosa [42] extracted fbank features as acoustic embedding; and for text modality, we follow the setting of the BART model [23].

4.6 Pre-training Tasks

To improve the emotional perception ability of the model, we propose novel pre-training tasks and divide the pre-training into two stages to guide the model to learn general emotional knowledge gradually.

4.6.1 Mask Context Modeling. We extend the Masked Language Modeling [5] to a multimodal input and context mask for the current query, named Mask Context Modeling (MCM).MCM can randomly mask tokens from the input text, acoustic, visual, and context modalities to encourage the model to learn to predict the masked tokens. To promote better generalization, we increase the mask probability to 50%. We denote the mask indices by $1 \leq m \leq M$, where M is the number of masked tokens. We denote the masked token by w_m, and the remaining tokens that are not masked by w_n. The loss function for MCM is defined as:

$$\mathcal{L}_{MCM}(\theta) = -\sum_{m=1}^{M} \log\left(P_\theta(w_m|w_n)\right), \tag{2}$$

where P_θ denotes the output distribution of the model, and θ represents model parameters to be optimized.

4.6.2 Sentiment Polarity Prediction. To encourage the model to learn to distinguish between different sentiment categories, we transform the fine-grained emotion labels of the datasets into sentiment polarity labels by mapping them to positive, negative, and neutral categories. Then, in the Sentiment Polarity Prediction (SPP) task, we train the model to predict the sentiment polarity category of the input. The loss function for SPP is defined as:

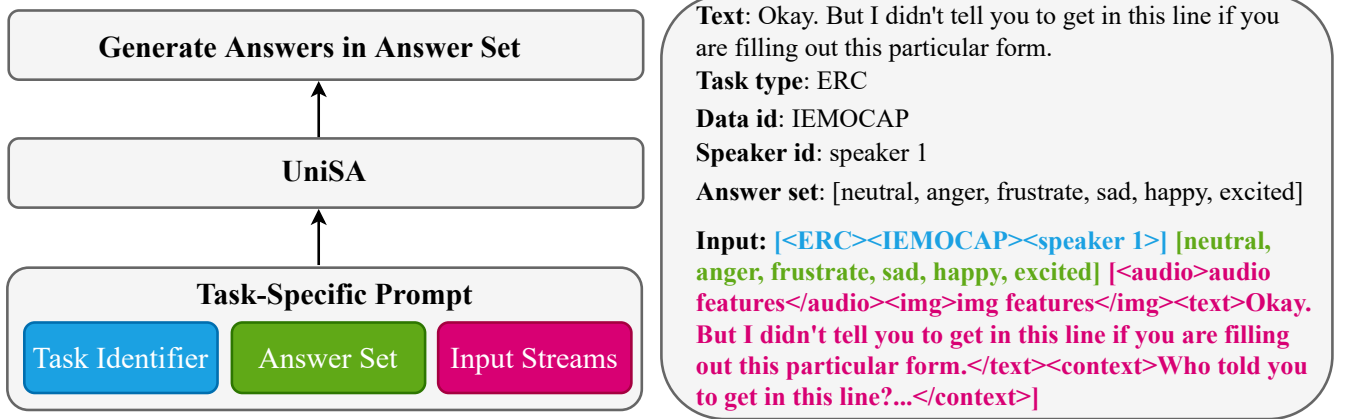$$\mathcal{L}_{SPP}(\theta) = -\log\left(P_\theta(s|w_n)\right). \tag{3}$$

Figure 3: An example of our Task-Specifical Prompt method.

4.6.3 Coarse-grained Label Contrast Learning. To encourage the model to learn to distinguish between different sentiment categories at a coarse level, we propose the Coarse-grained Label Contrast Learning (CCL) task. In this task, we take the output of the encoder as a representation of each sample and compute the Euclidean distance between samples with the same sentiment label in a batch. We then maximize the similarity between samples with the same sentiment label. The loss function for CCL is defined as:

$$\mathcal{L}_{CCL}(\theta) = \sum_{j=1}^{b} \frac{\sum_{k=1}^{b} (distance(j,k) * mask(j,k))}{\sum_{k=1}^{b} distance(j,k)}, \quad (4)$$

where b denotes the batch size, $distance(j,k)$ denotes the Euclidean distance between sample j and sample k, and $mask(j,k)$ is 1 when sample j and sample k have the same sentiment label, and 0 otherwise.

4.6.4 Cross-task Emotion Prediction. In the Cross-task Emotion Prediction (CEP) task, we first take the output of the encoder as the representation of each sample and cluster the samples for each subtask. Then, for each sample, we calculate the distance between its representation and each label cluster of each subtask. We take the label corresponding to the cluster with the smallest distance as the pseudo label of the sample for each subtask. Given a subtask set $D = \{ABSA, MSA, ERC, CA\}$, each sample will have four labels: one for the original label and three for the cross-task pseudo labels. The loss function for CEP is defined as:

$$\mathcal{L}_{CEP}(\theta) = -\sum^{D} \log (P_\theta(E_d|w_m)), \quad (5)$$

where E_d denotes the emotion label in subtask $d \in D$.

4.6.5 Pre-training Stage One. The first stage is coarse-grained emotion perception pre-training, including Mask Context Modeling, Sentiment Polarity Prediction, and Coarse-grained Label Contrast Learning. It aims to allow the model to acquire preliminary sentiment classification capabilities. Inspired by the Directional Expectation Test [55], we argue that emotions are polarity-invariant: combining queries with the same sentiment polarity does not change both. Therefore, we divide all datasets into different data pools

according to sentiment polarity. Then, we randomly select two queries from the same data pool to combine into a new query for the pre-training stage 1. The loss function for pre-training stage one is defined as:

$$\mathcal{L}_{stage1} = \mathcal{L}_{MCM} + \mathcal{L}_{SPP} + \mathcal{L}_{CCL}. \quad (6)$$

4.6.6 Pre-training Stage Two. The second stage is fine-grained emotion perception pre-training, including Mask Context Modeling and Cross-task Emotion Prediction. It aims to allow the model to acquire fine-grained emotion classification capability. The loss function for pre-training stage 2 is defined as:

$$\mathcal{L}_{stage2} = \mathcal{L}_{MCM} + \mathcal{L}_{CEP}. \quad (7)$$

5 EXPERIMENTS

5.1 Baseline

In this section, we reported the current SOTA models for each dataset in the SAEval benchmark, which we use as baselines to compare with the performance of UniSA on each subtask. These SOTA models are described as follows:

- **UniMSE** [18]: UniMSE is a multimodal sentiment knowledge-sharing framework that unifies MSA and ERC tasks from features, labels, and models. It is the current SOTA model for **MOSI** [77] and **MOSEI** [78].
- **EmoCaps** [30]: EmoCaps is a multimodal framework for conversational emotion recognition, which proposes the multimodal emotion vector to characterize the emotional tendencies of the utterance itself. It is the current SOTA model for **IEMOCAP** [4].
- **SPCL-CL-ERC** [62]: SPCL-CL-ERC is a model that combines Supervised Prototypical Contrastive Learning and curriculum learning strategy to address imbalanced classification problem in conversational emotion recognition. It is the current SOTA model for **EmoryNLP** [79] and **MELD** [50].
- **CoMPM** [22]: CoMPM is a model for conversational emotion recognition, which combines the speaker's pre-trained memory with the context model and finds that the pre-trained

Table 3: Experimental results on the SAEval benchmark. ★ indicates the model with the pre-training phase.

Task	Dataset	Metric	SOTA Models	SOTA Scores	$UniSA_{GPT2}$	$UniSA_{T5}$	$UniSA_{BART}$	$UniSA^{\star}_{BART}$
ABSA	SemEval14	WA ↑	InstructABSA	88.37	-	-	73.61	82.61
	SemEval16	WA ↑	InstructABSA	94.02	-	-	76.15	80.35
MSA	MOSI	MAE ↓		0.691	1.41	0.9004	0.9989	0.7422
		ACC-7 ↑	UniMSE	48.68	15.45	37.46	38.92	48.54
		ACC-2 ↑		85.85	44.75	76.82	77.11	84.11
	MOSEI	MAE ↓		0.523	0.8384	0.5460	0.5720	0.5866
		ACC-7 ↑	UniMSE	54.39	41.36	52.50	50.91	50.03
		ACC-2 ↑		85.86	71.02	84.22	85.57	84.93
ERC	MELD	WA ↑	SPCL-CL-ERC	-	48.12	64.52	62.45	62.34
		WF1 ↑		67.25	31.26	62.17	60.78	62.22
	IEMOCAP	WA ↑	EmoCaps	70.56	23.67	62.51	65.04	64.24
		WF1 ↑		71.77	9.06	62.70	65.21	64.46
	EmoryNLP	WF1 ↑	SPCL-CL-ERC	40.94	10.93	33.48	32.93	34.95
	DailyDialog	MF1 ↑	CoMPM	60.34	-	59.38	58.36	59.24
	EmoWoz	WF1 ↑	ContextBert	79.7	57.19	87.70	90.33	90.52
CA	SST-2	WA ↑	T5-11B	97.5	-	-	91.85	90.71
	IMDB	WA ↑	XLNet	96.21	-	-	93.35	92.26

Table 4: Results of ablation experiments on MOSI, MOSEI, IEMOCAP, and MELD.

	MOSI			MOSEI			MELD		IEMOCAP	
	MAE ↓	ACC-7 ↑	ACC-2 ↑	MAE ↓	ACC-7 ↑	ACC-2 ↑	WA ↑	WF1 ↑	WA ↑	WF1 ↑
SOTA Scores	0.691	48.68	85.85	0.523	54.39	85.86	67.85	66.71	70.56	71.77
S	1.79	20.69	44.75	1.006	41.31	71.02	46.59	31.37	20.22	60.27
S+P	1.06	28.86	69.67	0.6226	47.82	84.71	61.72	59.49	60.11	59.64
S+P+Text only	0.9192	41.39	78.86	0.6232	47.77	82.20	63.90	62.32	56.47	53.06
S+P+F	0.9439	29.59	74.63	0.5712	52.58	84.88	66.20	64.19	63.74	63.60
S+P+F+C	0.9034	41.25	78.13	0.5706	52.65	85.85	63.83	63.12	57.45	56.68
E+P+F+C	0.8908	40.52	77.55	0.5716	52.17	85.16	63.98	63.45	54.87	53.68
S+P+F+T1	0.8527	42.85	78.42	0.5544	53.12	85.81	62.60	61.03	57.15	56.97
S+P+F+T2	0.9441	37.17	78.71	0.5436	52.62	85.59	62.64	61.14	65.10	64.67
S+P+F+T1+T2	0.7422	48.54	84.11	0.5866	50.03	84.93	62.34	62.22	64.24	64.46

memory significantly improves the performance of the context model. It is the current SOTA model for **DailyDialog** [29].

In addition, the pre-trained models fine-tuned on specific datasets have achieved promising results: **BERT** [5] for **EmoWoz** [7]; **T5** [54] for **SST-2** [61]; **XLNet** [72] for **IMDB** [38]; **InstructABSA** [72] for **SemEval-2014** [48] and **SemEval-2016** [47].

5.2 Implementation Details

We explored three generative architectures, GPT-2 [53], T5 [54], and BART [23], as the backbone of our UniSA, and determined the optimal model through comparative experiments. The acoustic and visual representations have a hidden dimension of 64, while the textual embedding size is 768. The batch size is set to 64, and the learning rate is 5e-6 for BART-base, while it is 5e-5 for T5-base and GPT2-medium. Further details can be found in Table 2.

To evaluate the performance of the model on the primary tasks of sentiment analysis, we took the pre-trained UniSA, which undergoes two pre-training phases and conducted joint fine-tuning on the datasets in SAEval. During multi-task joint fine-tuning, the model may overfit on some tasks and underfit on others due to the varying learning difficulties across subtasks and inconsistent gradient directions between subtasks. To mitigate this issue, we proposed **task-average sampling** during the fine-tuning phase, which enables the model to learn the gradient of each task equally during each iteration. Specifically, we split the dataset into various task pools based on the task type, shuffled them, and distributed the same number of samples from each task pool equally for each step.

5.3 Performance Analysis

The experimental results of our UniSA and the existing SOTA models on the SAEval benchmark are presented in Table 3, where vacant cells indicate that the models were not fine-tuned on the corresponding datasets. We initially fine-tuned all datasets without pre-training stages using GPT2, T5, and BART as backbones. The results show that GPT-2 with a multimodal encoder ($UniSA_{GPT2}$) has limited performance. Furthermore, we observed that BART ($UniSA_{BART}$)

Table 5: Experimental results on the downstream tasks in terms of MF1. The SOTA models learn on the entire training set, while T5-base and UniSA are under the few-shot setting.

	Hateval	Twitter emoji	Twitter emotion	Twitter sentiment	Sarcasmania
Samples (few-shot/train)	300/9,000	3,000/4,5000	597/3,257	450/45,615	300/27,846
SOTA (train)	65.10	32.20	76.1	72.07	-
T5 (few-shot)	47.94	7.58	59.56	66.13	99.31
UniSA (few-shot)	49.61	14.82	65.98	64.17	99.78

outperforms T5 ($UniSA_{T5}$) in extending to multimodal architectures. **Therefore, we chose pre-trained BART as the backbone for our UniSA model in further experiments.**

As shown in Table 3, our proposed UniSA model performs comparably to the existing SOTA models for each dataset. Although UniSA is unable to outperform existing SOTA models on various benchmark datasets, it demonstrates the feasibility of uniformly modeling all sentiment analysis subtasks. Moreover, constrained by specific tasks and modalities, these SOTA models cannot be effectively generalized to other subtasks, whereas UniSA is an all-in-one model that can perform all sentiment analysis subtasks with a relatively small number of parameters.

5.4 Ablation Study

We used BART as the backbone of UniSA and conducted a series of ablation studies on the MOSI, MOSEI, IEMOCAP, and MELD datasets. Our ablation experiments aimed to investigate the effectiveness of our proposed methods and are as follows: 1) We replaced the proposed task specific prompt method with task tokens to verify its effects on model performance. 2) We removed the modal mask training method to verify its impact on model performance. 3) To explore the impact of different input forms on model performance, we introduced three additional LSTMs as encoders for audio, image, and context inputs. 4) We validated the effectiveness of each pre-training stage. 5) We eliminated the acoustic and visual modalities from the multimodal signals to investigate their effects on model performance.

The experimental results are shown in Table 4, where S denotes the sequences of all modalities input to a single encoder, while E denotes extended LSTMs as encoders for acoustic and visual modalities. P denotes Task-Specific Prompt, F denotes Modal Mask Training, C denotes additional LSTMs as an encoder for context, and T1 and T2 denote the first and second stages of pre-training, respectively. The results of the ablation experiments demonstrate the effectiveness of our proposed methods.

5.5 Few-shot Generalization

To demonstrate the generalizability of our proposed UniSA on various subtasks, we conducted experiments on several downstream tasks under the few-shot setting. We reported the SOTA scores of several downstream datasets, such as the SemEval2019 Hateval challenge [3], Sarcasmania [21], Semeval2017 Sentiment Analysis Challenge [56], Semeval2018 Emotion Recognition [43], Semeval2018 Emoji Prediction challenge [2], and performed few-shot experiments with a criterion of 150 samples/category.

The experimental results in Table 5 demonstrate that our UniSA, under low-resource settings, performs well on each downstream

dataset. The results reveal that UniSA has learned emotional knowledge common across subtasks during the pre-training stages and thus shows good generalization on various sentiment analysis subtasks.

5.6 Limitation Discussion

Error analysis in Appendix [4] reveals that the subjective bias among datasets is one of the factors limiting UniSA's performance. However, our experiments also suggest that there are other reasons for UniSA's limitations. One of these reasons is the performance limitations of the backbone models, which can significantly improve with more parameters and learned data [67]. Additionally, the lack of multimodal datasets poses another challenge for UniSA's performance. To address this, we encourage researchers to add more baseline datasets into SAEval and promote the development of multi-task unified modeling for sentiment analysis.

6 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new benchmark for Sentiment Analysis Evaluation, SAEval, and developed a multimodal generative framework named UniSA to recouple various subtasks of sentiment analysis. To overcome the challenges of unifying multi-tasks, we have introduced the Task Specific Prompt method and proposed novel pre-training tasks and training methods to improve the model's multimodal sentiment perception ability. Our extensive experiments have demonstrated the good generalizability of UniSA. We have also analyzed the bias between datasets and identified the limited performance of unified modeling for sentiment analysis subtasks.

In the future, we plan to expand our experiments by applying UniSA to more benchmark datasets and emotion-related tasks. We also intend to explore larger architectures as backbones for UniSA and introduce more pre-training data to improve its performance in various sentiment analysis tasks. Our work represents a step forward in the unified modeling of sentiment analysis subtasks, and we hope that it will inspire future research in this direction.

ACKNOWLEDGMENTS

---

[4]Please see Appendix A.2 for details.

REFERENCES

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *Advances in Neural Information Processing Systems* (2022), 32897–32912.

[2] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. 24–33.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 54–63.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42 (2008), 335–359.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* (2018).

[6] Rahul Dey and Fathi M Salem. 2017. Gate-Variants of Gated Recurrent Unit Neural Networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems*. 1597–1600.

[7] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 4096–4113.

[8] Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised Dialogue Topic Segmentation with Topic-aware Contrastive Learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2481–2485.

[9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2470–2481.

[10] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 154–164.

[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems* 30 (2017).

[12] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2594–2604.

[13] Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022. SPACE-2: Tree-Structured Semi-Supervised Contrastive Pre-training for Task-Oriented Dialog Understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*. 553–569.

[14] Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022. Unified Dialog Model Pre-Training for Task-Oriented Dialog Understanding and Generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 187–200.

[15] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Li Yongbin Si, Luo, et al. 2022. Galaxy: A Generative Pre-Trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10749–10757.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861* (2017).

[17] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. 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)*. 7042–7052.

[18] 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*. 7837–7851.

[19] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. 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)*. 5666–5675.

[20] Taewoon Kim and Piek Vossen. 2021. EmoBerta: Speaker-aware Emotion Recognition in Conversation with Roberta. *arXiv:2108.12009* (2021).

[21] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* (2023), 101861.

[22] Joosung Lee and Wooin Lee. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation.

[23] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[24] Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified Named Entity Recognition as Word-Word Relation Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10965–10973.

[25] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2022. GraphCFC: A Directed Graph based Cross-modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition. *arXiv:2207.12261* (2022).

[26] Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7056–7066.

[27] Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and Generation Make Bart a Good Dialogue Emotion Recognizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11002–11010.

[28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. 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)*. 2592–2607.

[29] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv:1710.03957* (2017).

[30] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*. 1610–1618.

[31] Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Duplex Conversation: Towards Human-like Interaction in Spoken Dialogue Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3299–3308.

[32] Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8360–8367.

[33] Jiangming Liu and Yue Zhang. 2017. Attention Modeling for Targeted Sentiment. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 572–577.

[34] 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:1907.11692* (2019).

[35] Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving Multimodal Sentiment Analysis via Multi-scale Fusion of Locally Descriptors. *arXiv:2112.01368* (2021).

[36] Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring Sequence-to-Sequence Learning in Aspect Term Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3538–3547.

[37] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4068–4074.

[38] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 142–150.

[39] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 164–172.

[40] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6818–6825.

[41] Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. DialogueTRM: Exploring the Intra-and Inter-modal Emotional

Behaviors in the Conversation. *arXiv:2010.07637* (2020).

[42] Brian McFee, Colin Raffel, Dawen Liang, Daniel P 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*. 18–25.

[43] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*. 1–17.

[44] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 188–197.

[45] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. 2022. Search-oriented micro-video captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3234–3243.

[46] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).

[47] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 Task 5: Aspect based Sentiment Analysis. In *ProWorkshop on Semantic Evaluation 2016*. 19–30.

[48] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation 2014)*. 27–35.

[49] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 873–883.

[50] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.

[51] Yushan Qian, Bo Wang, Ting-En Lin, Yinhe Zheng, Ying Zhu, Dongming Zhao, Yuexian Hou, Yuchuan Wu, and Yongbin Li. 2023. Empathetic Response Generation via Emotion Cause Transition Graph. *arXiv:2302.11787* (2023).

[52] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1113.

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI blog* (2019), 9.

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research* (2020), 5485–5551.

[55] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4902–4912.

[56] Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation 2017*. 502–518.

[57] Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. 2019. Revisiting LSTM Networks for Semi-supervised Text Classification via Mixed Objective Function. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6940–6948.

[58] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. DialogXL: All-in-one XLNet for Multi-party Conversation Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13789–13797.

[59] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. 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)*. 1551–1560.

[60] Shuzheng Si, Wentao Ma, Yuchuan Wu, Yinpei Dai, Haoyu Gao, Ting-En Lin, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue in Multiple Domains. *arXiv:2305.13040* (2023).

[61] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality over A Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1631–1642.

[62] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5197–5206.

[63] 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 57th Annual Meeting of the Association for Computational Linguistics*. 6558–6569.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in neural information processing systems* 30 (2017).

[65] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, et al. 2021. Ernie 3.0 Titan: Exploring Larger-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation. *arXiv:2112.12731* (2021).

[66] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 606–615.

[67] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *arXiv:2206.07682* (2022).

[68] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems* (2020), 6256–6268.

[69] Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation. 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)*. 525–535.

[70] Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A Unified Generative Framework for Aspect-based Sentiment Analysis. 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)*. 2416–2429.

[71] Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. 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)*. 5808–5822.

[72] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems* 32 (2019).

[73] Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-Text Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7900–7913.

[74] 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*. 10790–10797.

[75] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

[76] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[77] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv:1606.06259* (2016).

[78] 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)*. 2236–2246.

[79] Sayyed M Zahiri and Jinho D Choi. 2017. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. *arXiv:1708.04299* (2017).

[80] Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022. A Slot Is Not Built in One Utterance: Spoken Language Dialogs with Sub-Slots. In *Findings of the Association for Computational Linguistics: ACL 2022*. 309–321.

[81] Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. UniMS: A Unified Framework for Multimodal Summarization with Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11757–11764.

Table 6: Experimental results of subjective bias between datasets. ACC_{iem}, ACC_{meld}, ACC_{emo}, ACC_{mosi} **denote the accuracy under IEMOCAP, MELD, EmoryNLP, MOSI annotation systems, respectively.**

	ACC_{iem}	ACC_{meld}	ACC_{emo}	ACC_{mosi}
IEMOCAP	64.30	37.36	20.34	23.67
MELD	55.36	62.29	37.31	48.12
EmoryNLP	33.88	28.38	34.26	26.28
MOSI	31.48	23.90	31.63	48.54

A APPENDIX

A.1 Samples of SAEval Benchmark

In the SAEval Benchmark, all data is formatted into the dictionary, here are some examples:

```
{
    "task_type": ERC,
    "data_id": MELD,
    "speaker_id": speaker_1,      # Identify which speaker the current query belongs to
    "audio": [ audio_features ]   # audio features with 32x64 dim
    "image": [ image_features ]   # image features with 157x64 dim
    "text": also I was the point person on my company's transition from
            the KL-5 to GR-6 system.
    "context": <s>You must've had your hands full. <s> That I did. That I did.
    "utterance_id": 1             # Identify the current query's position in the conversation
    "label": happy
}
```

Figure 4: An example of the MELD data in SAEval.

```
{
    "task_type": CA,
    "data_id": IMDB,
    "speaker_id": no_one,          # Identifier for data that 'speaker_id' is unavailable
    "audio": no_audio_features     # Identifier for data that 'audio' is unavailable
    "image": no_image_features     # Identifier for data that 'image' is unavailable
    "text": One of the other reviewers has mentioned that after watching just 1 Oz
            episode you'll be hooked. The...
    "context": no_context          # Identifier for data that 'context' is unavailable
    "label": 1
}
```

Figure 5: An example of the IMDB data in SAEval.

Clik this link for more details: https://github.com/dawn0815/SAEval-Benchmark

A.2 Error Analysis

In this section, we further conducted experiments on several datasets to analyze the subjective bias of the datasets from both intra-task and inter-task perspectives.

We defined the annotation bias as the difference between the accuracy of a dataset under two different annotation systems, denoted as $Bias_{ana}$:

$$Bias_{ana} = |ACC_A - ACC_B|, \qquad (8)$$

where ACC_A denotes the accuracy under annotation system A and ACC_B refers to under accuracy under annotation system B. When there is no subjective bias between datasets D_i and D_j, the annotation bias of these two datasets should be approximately equal under both annotation systems, i.e., $Bias_{ana}^{(i)} \simeq Bias_{ana}^{(j)}$.

The subjective bias between datasets D_i and D_j can be formulated as:

$$Bias_{sub} = \left\| \left| ACC_A^{(i)} - ACC_B^{(i)} \right| - \left| ACC_A^{(j)} - ACC_B^{(j)} \right| \right\|, \qquad (9)$$

where $ACC_A^{(i)}$ denotes the accuracy of dataset D_i under annotation system A, and $ACC_B^{(j)}$ denotes the accuracy of dataset D_j under annotation system B.

To explore the existence of subjective bias, our experiment includes the following steps: 1) Obtain the representation of each sample in datasets $D_i \in \{IEMOCAP, MELD, MOSI, EmoryNLP\}$ from UniSA's encoder output. 2) Cluster samples in dataset D_i based on their real label. 3) For each sample s in D_i, calculate its distance to each cluster (category) in other datasets $D_j \in \{IEMOCAP, MELD, MOSI, EmoryNLP\}$, where $D_i \neq D_j$. 4) Select the cluster with the smallest distance as the pseudo-label of sample s under the annotation system D_j. 5) Generate emotion categories of dataset D_i under the labeling system of dataset D_j and calculate the accuracy through pseudo-labels.

The experimental results are shown in Table 6. According to Eqn. (9), we can calculate the subjective bias of IEMOCAP with MELD, EmoryNLP, and MOSI, which are 20.01%, 43.58%, and 23.57%, respectively. The subjective bias of MELD with EmoryNLP and MOSI is 19.1% and 10.47%, respectively, while the subjective bias of EmoryNLP with MOSI is 8.93%. These experimental results reveal intra-task subjective bias (IEMOCAP, MELD, and EmoryNLP) and inter-task subjective bias (MOSI with the three other datasets).

A.3 Training Details

All experiments are conducted with 8 x NVIDIA RTX V100 32G. The pre-training stage one took about 3 days, due to training millions of Amazon Reviews. The pre-training stage two and finetune took about 1 day, while the few-shot learning took only minutes to complete.