

Sentiment Knowledge Enhanced Self-supervised Learning for Multimodal Sentiment Analysis

Fan Qian, Jiqing Han[†], Yongjun He, Tieran Zheng, Guibin Zheng

School of Computer Science and Technology, Harbin Institute of Technology

{qianfan, jqhan, heyongjun, zhengtieran, zhengguibin}@hit.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) has made great progress that benefits from extraordinary fusion scheme. However, there is a lack of labeled data, resulting in severe overfitting and poor generalization for supervised models applied in this field. In this paper, we propose Sentiment Knowledge Enhanced Self-supervised Learning (SKESL) to capture common sentimental patterns in unlabeled videos, which facilitates further learning on limited labeled data. Specifically, with the help of sentiment knowledge and non-verbal behavior, SKESL conducts sentiment word masking and predicts fine-grained word sentiment intensity, so as to embed sentiment information at the word level into pre-trained multimodal representation. In addition, a non-verbal injection method is also proposed to integrate non-verbal information into the word semantics. Experiments on two standard benchmarks of MSA clearly show that SKESL significantly outperforms the baseline, and achieves new State-Of-The-Art (SOTA) results.

1 Introduction

Multimodal Sentiment Analysis (MSA) is a rapidly developing research field, which extends conventional text sentiment analysis to a multimodal setup where three modalities are present: text, audio and visual (Morency et al., 2011). With the abundance of user-generated opinion videos, MSA has a wide range of applications in e-commerce, intelligent customer service, human-computer interaction, etc.

In MSA, the construction of the training dataset relies on artificial perceptual evaluation for the sentiment of opinion videos, which is a very time-consuming and labor-intensive task, that is why video data with sentimental annotation is insufficient. As a result, supervised models applied in this field suffer from severe overfitting and poor

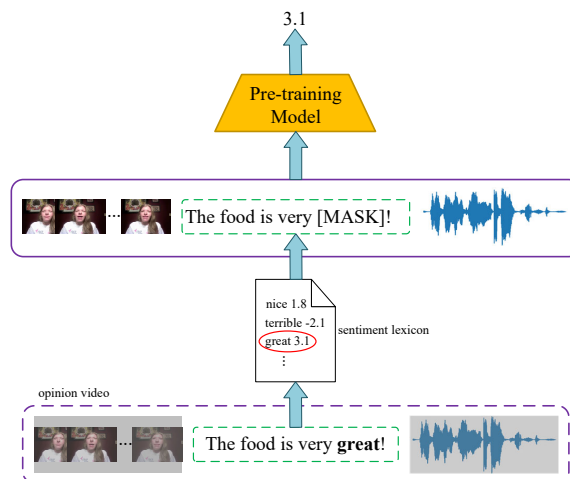


Figure 1: The pipeline of SKESL. The purple dashed box denotes an opinion video which includes text, visual and audio modalities. Visual and audio modalities with shaded boxes indicate not being used. The red circle represents the searched sentiment word.

generalization (Dai et al., 2021). Although previous studies have used several methods to alleviate the overfitting, most of them are based on general approaches such as multi-task learning (Dai et al., 2021; Akhtar et al., 2019; Chauhan et al., 2020; Yu et al., 2020), parameter regularization (Liang et al., 2019; Mai et al., 2020) and data augmentation (Liu et al., 2022), which neglect to consider the large number of unlabeled opinion videos that naturally exist on the Internet.

These opinion videos contain common sentimental patterns or compositional sentiment semantics about how the three modalities in the video are fused to express the overall sentiment, which can be leveraged to learn better sentiment representations. Inspired by recent knowledge-enhanced pre-training models on text sentiment analysis (Tian et al., 2020; Yin et al., 2020; Ke et al., 2020; Zhao et al., 2022), we argue that pre-training models enriched with the sentiment knowledge of words and non-verbal behavior will facilitate the characteriza-

[†] Corresponding author

tion of the sentimental patterns in videos, thereby resulting in better performance on multimodal sentiment analysis.

In this paper, we propose a Sentiment Knowledge Enhanced Self-supervised Learning (SKESL) method, which uses contextual and non-verbal information to predict the fine-grained sentiment intensity of a word to learn the common sentimental patterns in opinion videos, as shown in Figure 1. Specifically, given a speaker video without sentiment annotation, we first use the Automatic Speech Recognition (ASR) technology to obtain the transcribed text and then mask the most sentimentally salient words in the text according to the pre-specified sentiment lexicon. A pre-trained language representation model is utilized to acquire the sequence representation of the processed text. To integrate non-verbal information into the text representations, we further propose a non-verbal information aggregation method based on the cross-modal attention mechanism to derive non-verbal information-enhanced text representations. Finally, the masked word representations are exploited to predict the sentiment intensity itself.

After the SKESL is completed, we transfer the pre-trained model to the task of multimodal sentiment analysis, and adopt a small amount of sentiment-annotated data to fine-tune the model. To evaluate the effectiveness of SKESL, we test on two benchmark datasets: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh and Pu, 2018). Experimental results demonstrate that our model outperforms both the baseline and the current State-of-the-Art (SOTA) approach.

The main contributions can be summarized as follows:

- To the best of our knowledge, this paper is the first self-supervised learning method for multimodal sentiment analysis that leverages sentiment knowledge from large-scale unlabeled videos to facilitate improved sentiment representation learning.
- This paper proposes a novel non-verbal information aggregation method for obtaining text sequence representations enhanced by audio and visual information.
- The proposed SKESL method not only surpasses the baseline in experimental performance, but also achieves SOTA in the field of multimodal sentiment analysis.

2 Related Work

Pre-training Language Models In NLP, it has become a paradigm to pre-train language models on the large scale unlabeled data in an auto-encoding (Devlin et al., 2019) or auto-regressive manner (Radford et al., 2018, 2019), and fine-tune the pre-trained models on the downstream tasks using task-specific labeled data.

Recently, more pre-trained language models have been proposed, which can be roughly divided into four categories (Ke et al., 2020): 1) Knowledge enhancement: Introducing domain-specific knowledge in the process of pre-training language representation models has been shown to be effective. A representative model is ERNIE (Zhang et al., 2019), which explicitly introduces knowledge graph to pre-trained language models. 2) Transferability: On the basis of the general pre-training language model, further post-training is performed on more specific auxiliary tasks (Li et al., 2021). 3) Model compression: The compressed pre-trained language model can be widely applied in resource-constrained devices and tasks requiring real-time capability. The commonly used model compression methods include knowledge distillation (Sanh et al., 2019; Jiao et al., 2020), quantization (Shen et al., 2020) and pruning (Gordon et al., 2020). 4) Pre-training objectives: Aiming at rich and variable text expressions, many studies have further improved the text feature on basis of general Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objective (Devlin et al., 2019). For instance, SpanBERT (Joshi et al., 2020) masks consecutive spans randomly instead of individual tokens, while BERT-WWM (Cui et al., 2021) utilizes the Whole Word Mask (WWM) strategy to impose the model to learn complete semantics.

Knowledge Enhanced Pre-training Language Models Incorporating external knowledge into pre-training language models has become prevalent and has been shown to be significant. Such external knowledge includes commonsense knowledge for tasks such as entity typing and relation classification (Zhang et al., 2019; Peters et al., 2019; Liu et al., 2020; Xiong et al., 2020), sentiment knowledge for sentiment analysis (Tian et al., 2020; Yin et al., 2020; Ke et al., 2020), word sense knowledge for word sense disambiguation (Levine et al., 2020), commonsense knowledge for commonsense reasoning and sarcasm generation (Klein and Nabi, 2020; Chakrabarty et al., 2020), legal knowledge

for legal element extraction (Zhong et al., 2020), and biomedical knowledge for health question answering and medical inference (He et al., 2020).

Knowledge Enhanced Pre-training Models for Text Sentiment Analysis Some research (Tian et al., 2020; Yin et al., 2020; Ke et al., 2020; Zhao et al., 2022) integrates the sentiment knowledge into the pre-training process which includes sentiment words, word polarity and aspect-sentiment pairs. The learned representation would be more sentiment-specific and appropriate for *text* sentiment analysis.

Knowledge Enhanced Models for Multimodal Sentiment Analysis In MSA, some works consider sentiment knowledge with *explicit supervision*. For example, SWAFN (Chen and Li, 2020) designs a sentimental words prediction objective as an auxiliary task to incorporate sentimental words knowledge. MAGCN (Xiao et al., 2022) also incorporates sentiment knowledge into inter-modality learning.

3 Methodology

In this section, we describe our proposed Sentiment Knowledge Enhanced Self-supervised Learning (SKESL) framework for multimodal sentiment analysis, as shown in Figure 2. The framework contains Sentiment Word Masking (SWM), text representation learning, non-verbal information injection (a.k.a., multimodal fusion), and Sentiment Intensity Prediction (SIP) modules. In the subsequent subsections, we will detail the four modules.

3.1 Formulation

Our task is defined as follows: given a set of three modalities $M = \{T(\text{Text}), A(\text{Audio}), V(\text{Visual})\}$, an opinion video, i.e., multimodal sequence, can be represented as $\mathbf{X}^m = \{\mathbf{x}_1^m, \mathbf{x}_2^m, \dots, \mathbf{x}_{T_m}^m\} \in \mathbb{R}^{T_m \times d_m}$ where $m \in M - \{T\}$, $\mathbf{x}_i^m \in \mathbb{R}^{d_m}$ denotes the extracted sentiment feature corresponding to modality m , d_m is the dimension of the feature, and T_m is the length of sequence of modality m . Our goal is to predict the sentiment intensity $y \in \mathbb{R}$ or polarity $y \in \{\text{positive}, \text{neutral}, \text{negative}\}$ of the whole video.

3.2 Sentiment Word Masking

Sentiment Word Masking (SWM) aims to construct a corrupted version for each input sequence where sentiment information is masked. For a speaker video without sentiment annotation, a good ASR technique is first exploited to transcribe the speech

to text $S = \{w_1, w_2, \dots, w_N\}$. As sentiment words in the text, especially those with the most salient sentiment, are the most essential clues in the textual modality for detecting sentiment, we employ a sentiment lexicon to search and then mask them, i.e., use special tokens in their place. The sentiment lexicon (Hutto and Gilbert, 2014) consists of explicit sentiment intensity scores for each sentiment word, thus we can easily find the sentiment words with the highest sentiment intensity. Meanwhile, the score y_{MASK} of the highest sentiment intensity is chosen to act as a label for guiding SKESL. The corrupted sentence is represented as $S' = \{w_1, w_2, \dots, w_{\text{MASK}}, \dots, w_N\}$ where w_{MASK} denotes the masked word.

It is worth noting that a sentence with sentiment tendencies does not necessarily have sentiment words. To cope with this situation, we adopt a random masking strategy and assign a label with the sentiment intensity “0.0” to the masked word. The motivation is that the pre-training model is induced to distinguish whether the masked position holds a word without any sentiment based on contextual and non-verbal information. In this way, the model has a stronger sentimental semantic cognition of the words in the sentence and can learn better sentimental multimodal representations.

3.3 Text representation learning

After getting the corrupted sentence S' , we need to encode it into a sequence of word representations for subsequent processing. Given the outstanding language representation capabilities and widespread use of BERT, we opted to utilize it as the text encoder and input the corrupted sentence S' to it,

$$\mathbf{X}^T := \{\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_N^T\} = f_{\theta_{\text{LM}}}(S') \quad (1)$$

where θ_{LM} represents the parameters of BERT, $\mathbf{x}_i^T \in \mathbb{R}^{d_T}$ denotes the encoded word representation, and d_T is the dimension of the representation.

3.4 Non-verbal information injection

Unlike knowledge enhanced pre-training models towards text sentiment analysis, we emphasize that our SKESL deals with opinion videos that contain multiple modalities rather than just text. For the same word, there exists different sentiment with different non-verbal accompaniments. Therefore, the exact sentiment semantics of a word is determined by the word itself and the accompanied non-verbal behavior (Wang et al., 2019). Without the

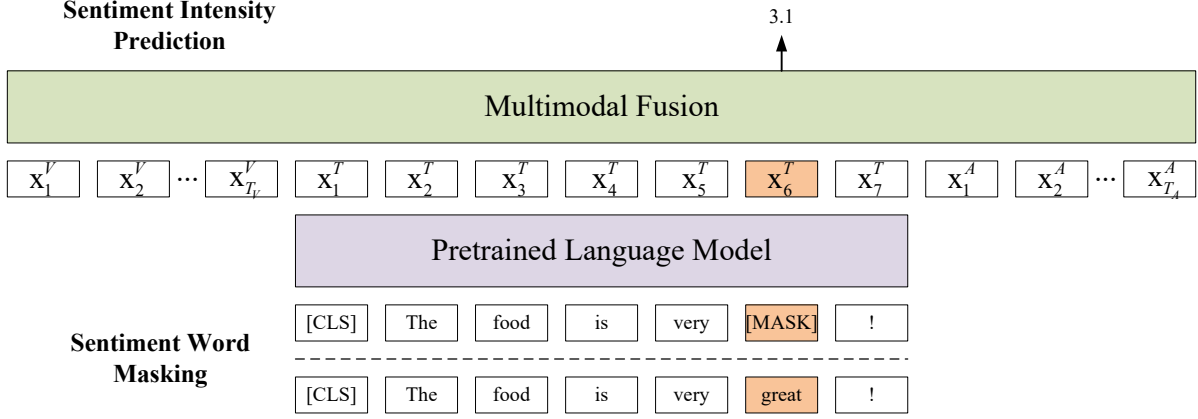


Figure 2: The model framework of Sentiment Knowledge Enhanced Self-supervised Learning (SKESL). SKESL contains two parts: (1) **Sentiment Word Masking** searches for the most sentimentally salient word of an input sentence based on the sentiment lexicon, and generates a corrupted version by replacing it with a special token [MASK]. (2) **Sentiment Intensity Prediction** requires the model to infer exact sentiment intensity according to contextual and non-verbal information.

help of non-verbal information, it is difficult for the pre-trained language model to determine the masked word and infer the sentiment intensity of the word. Therefore, to integrate the non-verbal information into word representations, inspired by Multimodal Transformer (MulT) (Tsai et al., 2019) which provides a latent cross-modal adaptation that fuses multimodal information by directly attending to low-level features in other modalities, we propose a new non-verbal information injection (a.k.a., multimodal fusion) method as shown in Figure 3.

The method repeatedly reinforces the text representations with the *low-level features* from audio and visual modalities by learning the attention across the features of two modalities. The low-level features benefit the model to preserve the original sentiment semantics for non-verbal behavior and learn the text-centric multimodal representations. Formally, we first define $\mathbf{X}_0^T = \mathbf{X}^T$, $\mathbf{X}_0^V = \mathbf{X}^V$ and $\mathbf{X}_0^A = \mathbf{X}^A$ to represent the text, visual and audio feature sequences before multimodal fusion, respectively. The Queries, Keys and Values sequences for Cross-Modal Attention (CMA) is computed by linear transformation as follows,

$$\mathbf{Q}^{mT} = \text{LN}(\mathbf{X}_{l-1}^T) \cdot \mathbf{W}_Q^m \quad (2)$$

$$\mathbf{K}^m = \text{LN}(\mathbf{X}_0^m) \cdot \mathbf{W}_K^m \quad (3)$$

$$\mathbf{V}^m = \text{LN}(\mathbf{X}_0^m) \cdot \mathbf{W}_V^m \quad (4)$$

where $m \in M - \{T\}$, $\text{LN}(\cdot)$ denotes the Layer Normalization, $\theta_{\text{CMA}} = \{\mathbf{W}_Q^m \in \mathbb{R}^{d_T \times d_T}, \mathbf{W}_K^m \in \mathbb{R}^{d_m \times d_T}, \mathbf{W}_V^m \in \mathbb{R}^{d_m \times d_T}\}$ are weights.

After obtaining Queries, Keys and Values sequences, we utilize the CMA to inject audio and visual information into text representations,

$$\begin{aligned} \mathbf{Y}_l^m &= \text{CMA}(\mathbf{Q}^{mT}, \mathbf{K}^m, \mathbf{V}^m) \\ &= \text{softmax}\left(\frac{\mathbf{Q}^{mT} \cdot \mathbf{K}^m}{\sqrt{d_T}}\right) \cdot \mathbf{V}^m \end{aligned} \quad (5)$$

where $m \in M - \{T\}$, $\mathbf{Y}_l^m \in \mathbb{R}^{N \times d_T}$ denotes the text sequence enhanced by modality m . In this way, each word receives information from all the elements across audio and visual feature sequences. Then, the enhanced text representations \mathbf{Y}_l^m along with the previous text representation \mathbf{X}_{l-1}^T are aggregated together,

$$\mathbf{Y}_l = \mathbf{Y}_l^A + \text{LN}(\mathbf{X}_{l-1}^T) + \mathbf{Y}_l^V \quad (6)$$

Similar to the structure of vanilla Transformer (Vaswani et al., 2017), $\mathbf{Y}_l \in \mathbb{R}^{N \times d_T}$ then goes through layer normalization, FeedForward Neural Network (FFNN) and residual connection,

$$\mathbf{X}_l^T = f_{\theta_{\text{FF}}}(\text{LN}(\mathbf{Y}_l)) + \mathbf{Y}_l \quad (7)$$

where θ_{FF} represents the parameters of FFNN, $\mathbf{X}_l^T \in \mathbb{R}^{N \times d_T}$ is the output of block l .

3.5 Sentiment Intensity Prediction

After L blocks, the refined text representations are $\mathbf{X}_L^T \in \mathbb{R}^{N \times d_T}$, in which $\mathbf{x}_{\text{MASK},L}^T$ is the refined representation of masked word. We simply use a 2-layer fully connected network with non-linear activation function to predict the sentiment intensity

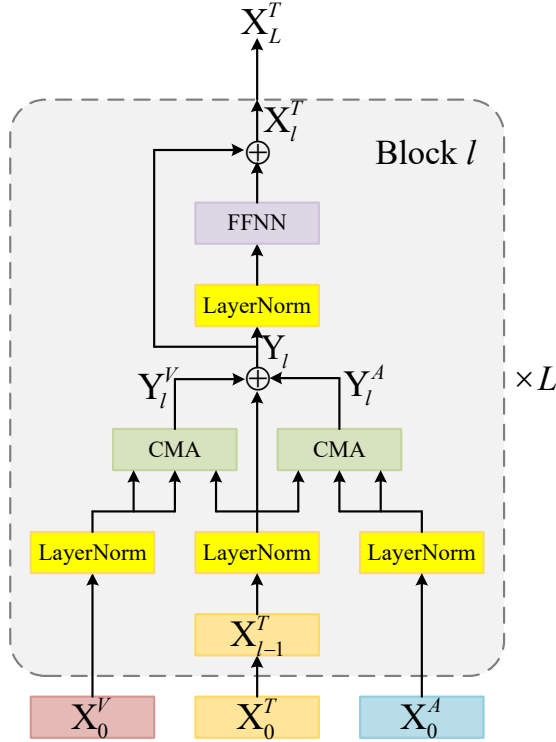


Figure 3: The framework of multimodal fusion. The superscripts $\{T, A, V\}$ denote text, audio and visual modalities, respectively.

of masked words.

$$y_{\text{pred}} = f_{\theta_{\text{FC}}}(\mathbf{x}_{\text{MASK},L}^T) \quad (8)$$

where θ_{FC} represents the parameters of the fully connected network, y_{pred} is the predicted sentiment intensity. We define $\theta = \{\theta_{\text{LM}}, \theta_{\text{CMA}}, \theta_{\text{FF}}, \theta_{\text{FC}}\}$, therefore the objective of the model is as follows,

$$\theta^* = \arg \min_{\theta} \mathcal{L}(y_{\text{pred}}, y_{\text{MASK}}) \quad (9)$$

where \mathcal{L} is chosen as Mean Absolute Error (MAE) loss function. The model is pre-trained in an end-to-end way.

3.6 Fine-tuning

We verify the effectiveness of SKESL on multimodal sentiment analysis task. On top of the pre-trained language model and multimodal fusion module, an output layer is added to perform task-specific prediction. The neural network is then fine-tuned on labeled multimodal data.

4 Experiments

4.1 Datasets

We pre-train our models on two speaker video datasets: **VoxCeleb1** (Nagrani et al., 2017) and

VoxCeleb2 (Chung et al., 2018) due to rich sentimental information within two datasets (Albanie et al., 2018). VoxCeleb1 and VoxCeleb2 contain over 100,000 video clips for 1,200+ speakers and over 1 million video clips for 6,000+ speakers collected from open-source media, respectively. Both datasets are approximately gender balanced, with speakers spanning a wide range of different ethnicities, accents, professions and ages. After a rough screening, we removed video clips that were not in English, and selected 132,708 video clips for 1,105 speakers from VoxCeleb1 and 947,726 video clips for 5,256 speakers from VoxCeleb2.

In addition, two multimodal sentiment datasets are used for fine-tuning and testing: **CMU-MOSI** (Zadeh et al., 2016), and **CMU-MOSEI** (Zadeh and Pu, 2018). The CMU-MOSI and CMU-MOSEI datasets consist of 2,199 and 22,846 opinion video clips from YouTube movie reviews, respectively. Each video clip has been scored between -3 (strong negative) and $+3$ (strong positive). Following previous works (Tsai et al., 2019; Rahman et al., 2020; Qian et al., 2022), for CMU-MOSI dataset, we utilize 1,284 segments for training, 229 segments for validation, and 686 segments for testing. For CMU-MOSEI dataset, we use 16,326 segments for training, 1,861 segments for validation, and 4,659 segments for testing. Table 1 and 2 show the statistics of all datasets.

4.2 Sentiment Features

We extract sentiment-related features for non-verbal modalities.

Audio: The library librosa (McFee et al., 2015) is used to extract frame-level acoustic features which include 1-dimensional logarithmic fundamental frequency (log F0), 20-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) and 12-dimensional Constant-Q chromatogram (CQT). These features are related to emotions and tone of speech according to (Yu et al., 2020).

Visual: The MultiComp OpenFace2.0 toolkit (Baltrusaitis et al., 2018) is used to extract a set of visual features including 340-dimensional facial landmarks, 35-dimensional facial action units, 6-dimensional head pose and orientation, 40-dimensional rigid and non-rigid shape parameters, and 288-dimensional eye gaze ¹.

¹For more details, you can see <https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format>

Dataset	# Video clips	# Speakers
VoxCeleb1	132,708	1,105
VoxCeleb2	947,726	5,256

Table 1: Datasets statistics for pre-training

Dataset	# Train	# Validation	# Test
CMU-MOSI	1,284	229	686
CMU-MOSEI	16,326	1,861	4,659

Table 2: Datasets statistics for fine-tuning and testing

4.3 Experimental Design

Sentiment Lexicon We use the VADER² sentiment lexicon (Hutto and Gilbert, 2014) to search and mask sentiment words. The VADER sentiment lexicon is sensitive to both the polarity and the intensity of sentiments expressed in social media contexts. It contains rich sentiment words with explicit sentiment scores from -4 to $+4$. In order to be consistent with the CMU-MOSI and CMU-MOSEI datasets, we linearly scale the score to $[-3, +3]$.

ASR We utilize the widely recognized Google Cloud Speech API³ to acquire transcripts for the pretraining datasets. Considering that we do not have access to the actual transcripts, it is not possible to calculate the precise ASR word error rate. However, we can confidently state that improved ASR performance would result in better outcomes. This is because a low-performing ASR system may inaccurately identify sentiment words, potentially leading to flawed results.

Training Details All models are built on the Pytorch (Paszke et al., 2019) toolbox with the NVIDIA RTX 3090 GPUs. The Adam (Kingma and Ba, 2014) optimizer is adopted for both pre-training and fine-tuning. The initial learning rate is set to $5e-6$ for BERT and $1e-4$ for other parameters. The batch size is 32. The number of epoch is 200. All our experiments were done with the exact same random seed. The models use the designated validation set of CMU-MOSI and CMU-MOSEI for finding best hyper-parameters⁴. You can refer to Appendices A for more details.

²You can find it at <https://github.com/cjhutto/vaderSentiment>

³<https://cloud.google.com/speech-to-text>

⁴Our codes are publicly available at <https://github.com/qianfan1996/SKESL.git>

4.4 Evaluation Metrics

Following previous works (Tsai et al., 2019; Yu et al., 2021), we record our experimental results in two forms: classification and regression. For classification, we report the weighted F1 score and binary accuracy. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values denote better performance for all metrics.

4.5 Baseline Models

Our model does not require manual alignment of language words with visual and audio, since the unlabeled video data has no explicit word timestamps. We perform a comprehensive comparative study against SKESL by considering various baselines and state-of-the-art models in either aligned or unaligned settings as detailed below.

4.5.1 Aligned Setting

MARN (Zadeh et al., 2018b) models intra-modal and cross-modal interactions by designing the Long-short Time Hybrid Memory and Multi-attention Block. MFN (Zadeh et al., 2018a) focuses on continuously modeling the view-specific and cross-view interactions, and aggregating them through time with a Multi-view Gated Memory. RMFN (Liang et al., 2018) decomposes the modeling process into multi-stage fusion, with each stage specifically targeting a subset of multimodal signals. RAVEN (Wang et al., 2019) considers the fine-grained structure of non-verbal subword sequences, and dynamically adjusts the word representations based on these non-verbal cues. MCTN (Pham et al., 2019) learns joint representations by cyclically translating from source to target modalities while ensures robustness even in the presence of noisy or missing target modalities. MISA (Hazari et al., 2020) incorporates a combination of losses including distributional similarity, orthogonal loss, reconstruction loss and task prediction loss to learn both modality-invariant and modality-specific representations. 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.

4.5.2 Unaligned Setting

MuT (Tsai et al., 2019) employs a cross-modal attention mechanism that enables a latent cross-modal adaptation, merging multimodal information by directly attending to low-level features in

Dataset Metric	CMU-MOSI				CMU-MOSEI				Setting
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow	
MARN \diamond	77.1	77.0	0.968	0.625	-	-	-	-	Aligned
MFN \diamond	77.4	77.3	0.965	0.632	76.0	76.0	-	-	
RMFN \diamond	78.4	78.0	0.922	0.681	-	-	-	-	
RAVEN \diamond	78.0	76.6	0.915	0.691	79.1	79.5	0.614	0.662	
MCTN \diamond	79.3	79.1	0.909	0.676	79.8	80.6	0.609	0.670	
MISA \diamond	83.4	83.6	0.783	0.761	85.5	85.3	0.555	0.756	
MAG-BERT \otimes	84.30	84.30	0.731	0.789	85.23	85.08	0.539	0.753	
MuT*	80.45	80.47	0.892	0.667	81.02	80.98	0.605	0.670	Unaligned
PMR*	81.33	81.30	0.875	0.669	82.12	82.07	0.614	0.675	
LMR-CBT*	80.42	80.38	0.901	0.657	80.75	80.79	0.634	0.653	
Self-MM*	85.21	85.18	0.773	0.774	84.07	84.12	0.556	0.750	
Ours	86.77	86.82	0.720	0.826	86.25	86.25	0.532	0.804	

Table 3: Results for multimodal sentiment analysis on CMU-MOSI and CMU-MOSEI datasets. NOTE: The unit of Acc and F1 is %. \uparrow means higher is better, and \downarrow is the opposite. \diamond means the result is from (Hazarika et al., 2020); \otimes from (Yu et al., 2021). And * denotes the reimplementation with non-verbal sentiment feature mentioned in 4.2. Best results are highlighted in bold.

other modalities. PMR (Lv et al., 2021) introduces a message hub to explore three-way interactions across all involved modalities within the context of multimodal fusion in unaligned multimodal sequences. LMR-CBT (Fu et al., 2021) achieves complementary learning of different modalities by incorporating three effective components: local temporal learning, cross-modal feature fusion and global self-attention representations. Self-MM (Yu et al., 2021) designs a label generation module to obtain independent unimodal supervisions, effectively balancing the learning progress across different sub-tasks.

5 Results and Analysis

In this section, we make a detailed analysis and discussion about our experimental results.

5.1 Quantitative Results

As shown in Table 3, our model achieves state-of-the-art performance on all the metrics on both datasets. Specifically, consistently significant improvement is observed compared to the previous unaligned models. Even comparing with aligned models, our model still achieves competitive or better results.

To investigate the impact of the amount of the unlabeled video data, we pre-train on VoxCeleb1 (132K) and VoxCeleb2 (947K) datasets and results are shown as Table 4 and 5. It is evident that a larger amount of pre-training data leads to more

Dataset Metric	CMU-MOSI			
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow
VoxCeleb1	86.43	86.46	0.725	0.818
VoxCeleb2	86.77	86.82	0.720	0.826

Table 4: Results on CMU-MOSI dataset with different amounts of pre-training data.

Dataset Metric	CMU-MOSEI			
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow
VoxCeleb1	85.34	85.31	0.550	0.778
VoxCeleb2	86.25	86.25	0.532	0.804

Table 5: Results on CMU-MOSEI dataset with different amounts of pre-training data.

significant performance improvements. Furthermore, we find that the performance improvement on the CMU-MOSEI dataset is larger than that on the CMU-MOSI dataset. For example, the accuracy is relatively improved by 0.39% and 1.07%, respectively. The most likely reasons are that the CMU-MOSI dataset is too small and contains noisy labels. Therefore, we boldly guess that it will be difficult to improve the performance on the CMU-MOSI dataset in the future.

In addition, to study the effect of different sizes of backbone language models, we use *bert-base* and *bert-large* models with 110M parameters and 340M parameters, respectively. Results are shown in Table 6 and 7. We observed that the perfor-


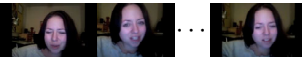



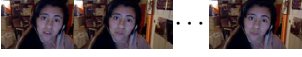
Audio	Text	Visual	Ground Truth	Ours	w/o SKESL
	Oh my gosh bad movie		-2.8	-2.6	-1.9
	I personally I liked Atlantis		1.6	1.3	2.2
	And I like how it shows		1.2	1.0	1.0

Figure 4: Examples from the CMU-MOSI dataset. For each example, we show the Ground Truth and prediction output of the model with and without SKESL.

Dataset Metric	CMU-MOSI			
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow
bert-base	86.02	85.97	0.734	0.815
bert-large	86.77	86.82	0.720	0.826

Table 6: Results on CMU-MOSI dataset under different pre-trained language models.

Dataset Metric	CMU-MOSEI			
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow
bert-base	85.32	85.27	0.553	0.762
bert-large	86.25	86.25	0.532	0.804

Table 7: Results on CMU-MOSEI dataset under different pre-trained language models.

mance improves as the pre-trained language model has more parameters and stronger expressiveness. This fits with our intuition and previous conclusion (Brown et al., 2020) that scaling up language models can greatly improve performance.

5.2 Ablation Study

To further explore the contributions of different components, we conduct an ablation study on CMU-MOSI dataset and the results are shown in Table 8. Without SKESL, the model’s accuracy and F1 score dropped by 1.40% and 1.46%, respectively. This suggests that it is indeed useful to transfer sentimental knowledge mined from unlabeled video data to downstream prediction tasks. Further, if the audio and visual modality is not used, i.e., only the BERT language model is used for the sentiment prediction task, the performance will be further degraded. This fact aligns with the observations in prior work (Tsai et al., 2019; Rahman et al., 2020; Qian et al., 2022) that multimodal sentiment

Dataset Metric	CMU-MOSI			
	Acc \uparrow	F1 \uparrow	MAE \downarrow	Corr \uparrow
Ours	86.77	86.82	0.720	0.826
w/o SKESL	85.37	85.36	0.726	0.815
w/o AV	85.06	85.12	0.735	0.808

Table 8: Ablation experiments on CMU-MOSI dataset. w/o denotes “without”. AV denotes audio & visual.

analysis is better than text-only sentiment analysis.

5.3 Case Study

To visually validate the reliability of our model, we present some examples shown in Figure 4. The examples are from the first three speaker videos of the test set of the CMU-MOSI dataset. The prediction results demonstrate that pre-training models enriched with the sentiment knowledge of words and non-verbal behavior will facilitate the characterization of the sentimental patterns in videos, thereby resulting in better performance on multi-modal sentiment analysis.

6 Conclusion

In this paper, we highlighted the sentiment knowledge enhanced self-supervised learning in MSA. We find that mining sentimental prior information from unlabeled video data can lead to better predictions on labeled data. The larger the amount of unlabeled video data and the stronger the language modeling ability, the better the performance that can be achieved.

7 Limitations

We note that there are several limitations with such a sentiment knowledge enhanced self-supervised

learning approach. First, the preprocessing of massive videos is time-consuming and laborious. Second, the pre-training of our model has relatively large requirements on the GPU resources. Finally, we argue that there should not be too many videos without sentimental words, so as to avoid the model having a large bias and not learning any sentiment knowledge.

References

- Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *ArXiv*, abs/1905.05812.
- Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *ArXiv*, abs/2004.13248.
- Dushyant Singh Chauhan, R DhanushS., Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *ACL*.
- Minping Chen and Xia Li. 2020. Swafn: Sentimental words aware fusion network for multimodal sentiment analysis. In *COLING*.
- Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Wenliang Dai, Samuel Cahyawijaya, Yejin Bang, and Pascale Fung. 2021. Weakly-supervised multi-task learning for multimodal affect recognition. *ArXiv*, abs/2104.11560.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- Ziwan Fu, Feng Liu, Hanyang Wang, Siyuan Shen, Jiahao Zhang, Jiayin Qi, Xiangling Fu, and Aimin Zhou. 2021. Lmr-cbt: Learning modality-fused representations with cb-transformer for multimodal emotion recognition from unaligned multimodal sequences. *arXiv preprint arXiv:2112.01697*.
- Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing bert: Studying the effects of weight pruning on transfer learning. In *REPLANLP*.
- 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.
- Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *EMNLP*.
- Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. *ArXiv*, abs/1909.10351.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *EMNLP*.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. *ArXiv*, abs/2005.00669.
- Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. Sensebert: Driving some sense into bert. *ArXiv*, abs/1908.05646.

- Zhongyang Li, Xiao Ding, and Ting Liu. 2021. Transbert: A three-stage pre-training technology for story-ending prediction. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20:16:1–16:20.
- 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. *ArXiv*, abs/1907.01011.
- Paul Pu Liang, Liu Ziyin, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *EMNLP*.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*.
- Yih-Ling Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. *ArXiv*, abs/2209.02604.
- Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562.
- Sijie Mai, Songlong Xing, Jia-Xuan He, Ying Zeng, and Haifeng Hu. 2020. Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. *ArXiv*, abs/2011.13572.
- 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*, volume 8, pages 18–25.
- Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.
- Arsha Nagrani, Joon Son Chung, and Andrew Senior. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703.
- Matthew E. Peters, Mark Neumann, IV Robert L. Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *EMNLP*.
- Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.
- Fan Qian, Hongwei Song, and Jiqing Han. 2022. Word-wise sparse attention for multimodal sentiment analysis. *Proc. Interspeech 2022*, pages 1973–1977.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *AAAI*.
- Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In *ACL*.
- 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.
- 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.
- Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

- Luwei Xiao, Xingjiao Wu, Wen Wu, Jing Yang, and Liangbo He. 2022. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4578–4582.
- Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *ArXiv*, abs/1912.09637.
- Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *ACL*.
- 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 *ACL*.
- Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *ArXiv*, abs/2102.04830.
- 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, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Amir Zadeh and Paul Pu. 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 (Long Papers)*.
- 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 preprint arXiv:1606.06259*.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *ACL*.
- Qinghua Zhao, Shuai Ma, and Shuo Ren. 2022. Kesa: A knowledge enhanced approach for sentiment analysis. *ArXiv*, abs/2202.12093.
- Haoxiang Zhong, Chaojun Xiao, Cunchao Tu, T. Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. In *ACL*.

A Appendices

A.1 Hyper-parameters Setting

Hyper-parameter	Value
d_A	33
d_V	709
d_T	768&1024
L	4
Dimension of fully connected layer	256
Batch size	32
Number of epoch	200
Learning rate of BERT	5e-6
Learning rate of other parameters	1e-4
Optimizer	Adam

Table 9: The hyper-parameters of the model.

A.2 Model Efficiency

	Computational Budget	# Parameters
<i>bert-base</i>	4.6 GMACs	111.4 M
<i>bert-large</i>	13.5 GMACs	337.9 M

Table 10: The efficiency of the models.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
section 7
- A2. Did you discuss any potential risks of your work?
Our work does not have potential risks.
- A3. Do the abstract and introduction summarize the paper’s main claims?
section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C Did you run computational experiments?

section 4.5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendices A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

section 4.3 and Appendices A.1

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

As with the original paper for other work (e.g., MAG-BERT, MulT, Self-MM) in MSA, we do not report error bars. All experiments were done with the exact same random seed. We report the max after a grid search of hyperparameter using the validation sets.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

section 4.2, 4.3

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.