Sentiment Word Aware Multimodal Refinement for Multimodal Sentiment Analysis with ASR Errors

Yang Wu[1] **Yanyan Zhao**[1*] **Hao Yang**[1] **Song Chen**[1] **Bing Qin**[1]
Xiaohuan Cao[2] **Wenting Zhao**[2]
[1] Harbin Institute of Technology [2] AI Lab of China Merchants Bank
[1] {ywu, yyzhao, hyang, songchen, qinb}@ir.hit.edu.cn
[2] {xhcao, wtzhao}@cmbchina.com

Abstract

Multimodal sentiment analysis has attracted increasing attention and lots of models have been proposed. However, the performance of the state-of-the-art models decreases sharply when they are deployed in the real world. We find that the main reason is that real-world applications can only access the text outputs by the automatic speech recognition (ASR) models, which may be with errors because of the limitation of model capacity. Through further analysis of the ASR outputs, we find that in some cases the sentiment words, the key sentiment elements in the textual modality, are recognized as other words, which makes the sentiment of the text change and hurts the performance of multimodal sentiment analysis models directly. To address this problem, we propose the sentiment word aware multimodal refinement model (SWRM), which can dynamically refine the erroneous sentiment words by leveraging multimodal sentiment clues. Specifically, we first use the sentiment word position detection module to obtain the most possible position of the sentiment word in the text and then utilize the multimodal sentiment word refinement module to dynamically refine the sentiment word embeddings. The refined embeddings are taken as the textual inputs of the multimodal feature fusion module to predict the sentiment labels. We conduct extensive experiments on the real-world datasets including MOSI-Speechbrain, MOSI-IBM, and MOSI-iFlytek and the results demonstrate the effectiveness of our model, which surpasses the current state-of-the-art models on three datasets. Furthermore, our approach can be adapted for other multimodal feature fusion models easily[1].

1 Introduction

Multimodal sentiment analysis (MSA) has been an emerging research field for its potential applications in human-computer interaction. How to effectively fuse multimodal information including textual, acoustic, and visual to predict the sentiment is a very challenging problem and has been addressed by many previous studies. Some works focus on introducing additional information into the fusing model, such as the alignment information between different modal features (Wu et al., 2021) and unimodal sentiment labels (Yu et al., 2021). And other works consider the semantic gaps between multimodal data and adopt the adversarial learning (Mai et al., 2020) and multi-task learning (Hazarika et al., 2020) to map different modal features into a shared subspace.

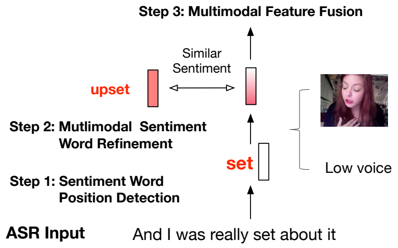
Datasets	MAE ↓
MOSI-Gold	72.49
MOSI-SpeechBrain	90.95(↑18.46)
MOSI-IBM	85.65(↑13.16)
MOSI-iFlytek	78.79(↑ 6.30)

(a) Results of the Self-MM model on the real-world datasets. SpeechBrain, IBM, and iFlytek are three ASR APIs we adopted.

Gold:
And I was really **upset** about it
ASR:
And I was really **set** about it

Datasets	Percentages
MOSI-SpeechBrain	26.5%
MOSI-IBM	17.6%
MOSI-iFlytek	10.6%

(b) An example of the sentiment word substitution error and the percentages of it on the datasets.


Step 3: Multimodal Feature Fusion
Similar Sentiment
upset
Step 2: Mutlimodal Sentiment Word Refinement
set
Low voice
Step 1: Sentiment Word Position Detection
ASR Input
And I was really set about it

(c) Our approach to reduce the negative impact of the sentiment word substitution error on the MSA models.

Figure 1: Illustration of our motivation.

Despite the apparent success of the current state-

[*] Corresponding Author
[1] Data and code are available at https://github.com/albertwy/SWRM

Findings of the Association for Computational Linguistics: ACL 2022, pages 1397 - 1406
May 22-27, 2022 ©2022 Association for Computational Linguistics

of-the-art models, their performance decreases sharply, when deployed in the real world. The reason is that the input texts are provided by the ASR models, which usually are with errors because of the limitation of model capacity. To further analyze this problem, we build three real-world multimodal sentiment analysis datasets based on the existing dataset, CMU-MOSI(Zadeh et al., 2016). Specifically, we adopt three widely used ASR APIs including SpeechBrain, IBM, and iFlytek to process the original audios and obtain the recognized texts. Then, we replace the gold texts in CMU-MOSI with the ASR results and get three real-world datasets, namely MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek. We evaluate the current state-of-the-art model, Self-MM(Yu et al., 2021), and report the mean absolute error (MAE) on the multimodal sentiment analysis task. As we can see in Figure 1(a), when the model is deployed in the real world, there is an obvious drop in model performance.

The further in-depth analysis of ASR errors shows that the sentiment word substitution error can hurt the MSA model directly. The reason is that the sentiment words in the text are the most important clues in the textual modality for detecting sentiment and incorrectly recognizing them could change the sentiment conveyed by the text. To have an intuitive understanding of the sentiment word substitution error, we take an example in Figure 1(b). The gold text is “And I was really upset about it", but the ASR model (SpeechBrain) recognizes the sentiment word “upset" wrongly as “set", which results in the change of the sentiment semantics of the text and directly affects the MSA model performance. We list the percentages of the sentiment word substitution error on the MOSI dataset for three ASR APIs in Figure 1(b). The percentage of the sentiment word substitution error on the MOSI-IBM is 17.6%, which means about 17 of 100 utterances have this type of error. To further demonstrate the negative effect of the substitution error on the MSA models, we split the test data of MOSI-IBM into two groups by whether there is a substitution error. We evaluate Self-MM on the test data and observe that the misclassification rate of the group in which the substitution error exists is higher than the other group (29.9% vs 15.8%). This result indicates that the sentiment word substitution error could hurt the state-of-the-art MSA model.

To tackle this problem, we propose the sentiment word aware multimodal refinement model, which can detect the positions of the sentiment words in the text and dynamically refine the word embeddings in the detected positions by incorporating multimodal clues. The basic idea of our approach is shown in Figure 1(c). We consider leveraging the multimodal sentiment information, namely the negative sentiment conveyed by the low voice and sad face, and textual context information to help the model reconstruct the sentiment semantics for the input embeddings. Specifically, we first use the sentiment word location module to detect the positions of sentiment words and meanwhile utilize the strong language model, BERT, to generate the candidate sentiment words. Then we propose the multimodal sentiment word refinement module to refine the word embeddings based on the multimodal context information. The refinement process consists of two parts, filtering and adding. We apply the multimodal gating network to filter out useless information from the input word embeddings in the filtering process and use the multimodal sentiment word attention network to leverage the useful information from candidate sentiment words as the supplement to the filtered word embeddings in the adding process. Finally, the refined sentiment word embeddings are used for multimodal feature fusion.

We conduct extensive experiments on the MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek datasets to demonstrate the effectiveness of our proposed model. The experimental results show that: (1) There is an obvious performance drop for the state-of-the-art MSA model, when the model is deployed in the real world taking the ASR outputs as the input of textual modality; (2) Our proposed model outperforms all baselines, which can dynamically refine the sentiment word embeddings by leveraging multimodal information.

The main contributions of this work are as follows: (1) We propose a novel sentiment word aware multimodal refinement model for multimodal sentiment analysis, which can dynamically reconstruct the sentiment semantics of the ASR texts with errors by utilizing the multimodal sentiment information resulting in more robust sentiment prediction; (2) We validate the negative effect of the sentiment word substitution error on the state-of-the-art MSA model through the in-depth analysis; (3) We evaluate our model on three real-world datasets, and the experimental results demonstrate that our model outperforms all baselines.

2 Related Work

Multimodal sentiment analysis has gained increasing attention from the community recently and some process has been made. In general, there are three findings presented by previous work.

Performing the cross-modal alignment is helpful for multimodal feature fusion. Chen et al. (2017) considered that the holistic features mainly contain global information, which may fail to capture local information. Therefore, they applied the force-alignment to align the visual and acoustic features with the words and further obtained the word-level features. To effectively fuse them, they proposed the GME-LSTM(A) model, which consists of two modules, the gated multimodal embedding and the LSTM with the temporal attention. However, obtaining the word-level features needs to perform the force-alignment, which is time-consuming. To address it, Tsai et al. (2019) proposed the MulT model, which uses the cross-modal attention to align different modal features implicitly. Instead of performing the alignment in the time dimension, some works focusing on semantic alignment. Hazarika et al. (2020) considered that the semantic gaps between heterogeneous data could hurt the model performance and proposed the MISA model, which maps the different modal data into a shared space before multimodal feature fusion. Wu et al. (2021) first utilized the cross-modal prediction task to distinguish the shared and private semantics of non-textual modalities compared to the textual modality and then fuse them. The above works show that performing the cross-modal alignment is helpful for multimodal feature fusion.

Training the MSA models in an end-to-end manner is more effective. Most of the previous studies adopt a two-phase pipeline, first extracting unimodal features and then fusing them. Dai et al. (2021) considered that it may lead to sub-optimal performance since the extracted unimodal features are fixed and cannot be further improved benefiting from the downstream supervisory signals. Therefore, they proposed the multimodal end-to-end sparse model, which can optimize the unimodal feature extraction and multimodal feature fusion jointly. The experimental results on the multimodal emotion detection task show that training the models in an end-to-end manner can obtain better results than the pipeline models.

Leveraging the unimodal sentiment labels to learn more informative unimodal representations is useful for multimodal feature fusion. Yu et al. (2020) considered that introducing the unimodal sentiment labels can help the model capture the unimodal sentiment information and model the difference between modalities. Motivated by it, they built the CH-SIMS dataset, which contains not only the multimodal sentiment labels but also unimodal sentiment labels. And based on it, they proposed a multi-task learning framework to leverage two types of sentiment labels simultaneously. However, this method needs unimodal labels, which is absent for most of the existing datasets. To address it, Yu et al. (2021) proposed the Self-MM model, which first generates the unimodal labels by utilizing the relationship between the unimodal and multimodal labels and then uses the multi-task learning to train the model. These two works both address the usefulness of introducing unimodal labels.

However, even though lots of models are proposed and obtain promising results on the benchmark datasets, there are few works considering the noisy inputs when the MSA models are deployed in the real world. Chen et al. (2017) presented the Gated Multimodal Embedding to filter out the noises from the acoustic and visual data. Pham et al. (2019) considered that visual and acoustic data may be absent and proposed the MCTN model to handle it. Liang et al. (2019) and Mittal et al. (2020) also mainly focused on dealing with the noises introduced by the visual and acoustic data, and their models are based on the word-level features, which are obtained by aligning the audios with the gold texts. There is only one work (Dumpala et al., 2018) considering that the texts are output by the ASR models, which may be erroneous. But this work does not study how do the ASR errors affect the MSA models and does not evaluate the SOTA MSA models on the datasets. Besides, the proposed model needs the gold texts when training, which is time-consuming and labor-consuming.

Comparing to the above works, we evaluate the SOTA MSA models on the real-world datasets and observe that the performance of models decreases sharply because of the erroneous ASR texts. Through in-depth analysis of the ASR outputs, we find the sentiment word substitution error in the ASR texts could hurt the MSA models directly. To address it, we propose the sentiment word aware multimodal refinement model, which only uses the

ASR texts in the training and testing phrases.

3 Approach

In this section, we describe the sentiment word aware multimodal refinement model in detail. An illustration of our proposed model is given in Figure 2. Our model consists of three modules including the sentiment word location module, multimodal sentiment word refinement module, and multimodal feature fusion module. We first use the sentiment word location module to detect the possible positions of sentiment words and then utilize the multimodal sentiment word refinement module to dynamically refine the word embeddings in the detected positions. Finally, the refined word embeddings are fed into the multimodal feature fusion module to predict the final sentiment labels.

3.1 Sentiment Word Position Detection

The core idea of the sentiment word position detection module is to find out the possible positions of sentiment words in the ASR texts. Note that, it is different from locating sentiment words depending on the word semantics, since the ASR models may recognize a sentiment word as a neutral word, which makes it hard to locate correctly. For example, given a gold text “And I was really upset about it", the ASR model recognizes it as “And I was really set about it". It is easy for the model to label the word “set" as a neutral word. Therefore, we choose to detect the position of the sentiment words instead of locating them.

To achieve it, we consider adopting a powerful language model, since the language model can model the context information of the sentiment words such as syntactic and grammatical information and predict the appropriate words for the target position. Specifically, we choose the BERT model (Devlin et al., 2019) as our language model since the masked language modeling pretraining objective meets our needs perfectly. Given the sentence $\{w_1, w_2, ..., w_{n_l}\}$, we first mask each word w_i in the sentence sequentially, and in practice, we replace the word with the special word [MASK]. For example, we mask the first word in the sentence and obtain $\{[\mathrm{MASK}], w_2, ..., w_{n_l}\}$. And then we use the BERT model to predict the possible words in the position of the masked word. We sort the predicted candidate words by the prediction probabilities and get the Top-k candidate words $C_i = \{c_1^i, c_2^i, ..., c_k^i\}$.

Next, we distinguish the sentiment words from the candidates using the sentiment lexicons (Hu and Liu, 2004; Wilson et al., 2005) and k_i is the number of selected sentiment words corresponding to the position i. The larger the number is, the more possible the position is. And we obtain the most possible position of sentiment word, $s = \arg\max(\{k_1, k_2, ..., k_{n_l}\})$. Considering that in some cases there is not a sentiment word in the sentence, we use a sentiment threshold to filter out the impossible ones. In practice, we use the gate mask p to record it, and p is 1 if k_s is larger than $k/2$ and 0 otherwise.

3.2 Multimodal Sentiment Word Refinement

In order to reduce the negative effects of the ASR errors, we propose the multimodal sentiment word refinement module, in which we refine the word embeddings of sentiment words from two aspects. One is that we uses the multimodal gating network to filter out the useless information from the input word embeddings. The other one is that we design the multimodal sentiment attention network to incorporate the useful information from candidate words generated by the BERT model.

Given an utterance, which includes three modal unaligned features, word embeddings, acoustic features, and visual features, we denote them as $\mathbf{x}^i = \{x_t^i : 1 \leq t \leq n_i, x_t^i \in \mathbb{R}^{d_x^i}\}$, $i \in \{l, v, a\}$. To obtain the multimodal information corresponding to each word, We utilize the pseudo-alignment method to align the features. We split the the acoustic and visual features into non-overlapping feature groups, of which lengths are $\left\lfloor \frac{n_a}{n_l} \right\rfloor$ and $\left\lfloor \frac{n_v}{n_l} \right\rfloor$ respectively, and average the features in each group and obtain the aligned features, $\mathbf{u}^i = \{u_t^i : 1 \leq t \leq n_l, u_t^i \in \mathbb{R}^{d_x^i}\}$, $i \in \{v, a\}$.

To obtain the context-aware representations, we apply the BERT model and LSTM networks to encode the features, producing $\mathbf{h}^i = \{h_t^i : 1 \leq t \leq n_l, h_t^i \in \mathbb{R}^{d_h^i}\}$, $i \in \{v, a, l\}$. Besides, we also use an LSTM network to fuse the acoustic and visual features for capturing high-level sentiment semantics and obtain $\mathbf{h}^{va} = \{h_t^{va} : 1 \leq t \leq n_l, h_t^{va} \in \mathbb{R}^{d_h^{va}}\}$.

$$
\begin{aligned}
\mathbf{h}^l &= \mathrm{BERT}(\mathbf{x}^{\mathrm{l}}) \\
\mathbf{h}^v &= \mathrm{LSTM}_{\mathrm{v}}(\mathbf{u}^{\mathrm{v}}) \\
\mathbf{h}^a &= \mathrm{LSTM}_{\mathrm{a}}(\mathbf{u}^{\mathrm{a}}) \\
\mathbf{h}^{va} &= \mathrm{LSTM}_{\mathrm{va}}([\mathbf{u}^{\mathrm{v}}; \mathbf{u}^{\mathrm{a}}])
\end{aligned} \tag{1}
$$

Subsequently, We propose the multimodal gating

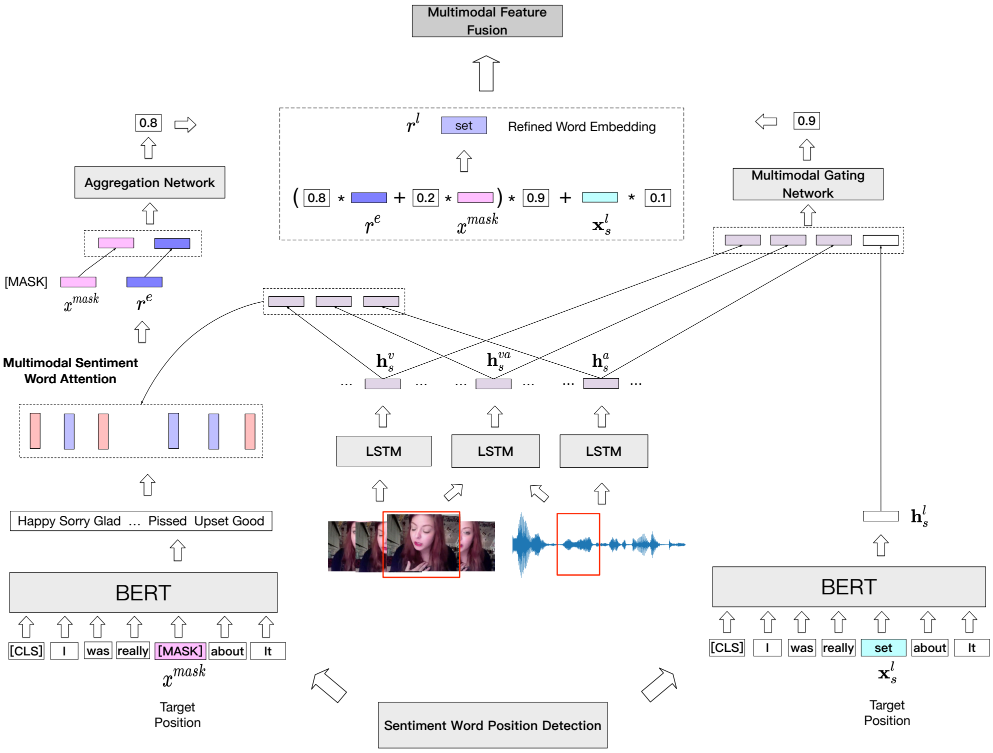


Figure 2: Illustration of our proposed model.

.

network to filter the word embedding, which is implemented by a non-linear layer. The motivation is that the ASR model may recognize incorrectly the sentiment word resulting in the corrupted sentiment semantics of the text. Therefore, we leverage the multimodal sentiment information to decide how much information of the input word embedding to pass. Specifically, we concatenate the unimodal context-aware representations, $\mathbf{h}_s^l$, $\mathbf{h}_s^v$, $\mathbf{h}_s^a$, and bimodal representation $\mathbf{h}_s^{va}$ in the detected position s and feed them into a non-linear neural network, producing the gate value g^v. And then the gate value is used to filter out the useless information from the word embedding. To make the model ignore the impossible one, we use the gate mask p to achieve it.

$$\begin{aligned} g^v &= Sigmoid(W_1([\mathbf{h}_s^l; \mathbf{h}_s^v; \mathbf{h}_s^a; \mathbf{h}_s^{va}]) + b_1) \\ r^v &= (1 - g^v p)\mathbf{x}_s^l \end{aligned} \tag{2}$$

where $W_1 \in \mathbb{R}^{1\times\sum_{i\in\{l,v,a,va\}} d_h^i}$, $b_1 \in \mathbb{R}^1$ are the parameters of the multimodal gating network.

Furthermore, we propose a novel multimodal sentiment word attention network to leverage the sentiment-related information from the candidate words, more than half of which are sentiment words, generated by the BERT model to complement the word embeddings. For example, the ASR model recognizes the “upset" as “set", we first want to remove the useless information of “set" and then incorporate the information of negative sentiment words to reconstruct the original sentiment semantics. We use a linear layer to implement the multimodal sentiment word attention network. We first concatenate the word embedding $x^{c_t^s}$ of the candidate word c_t^s and multimodal representations, $\mathbf{h}_s^v$, $\mathbf{h}_s^a$, and $\mathbf{h}_s^{va}$ at the most possible time step s. Then, we pass them to the linear layer and obtain the attention score g_t^e. The attention scores are fed into a softmax function to obtain the attention weights. Finally, we apply the weights to the candidate word embeddings and get the sentiment embedding r^e.

$$\begin{aligned} g_t^e &= W_2([x^{c_t^s}; \mathbf{h}_s^v; \mathbf{h}_s^a; \mathbf{h}_s^{va}]) + b_2 \\ w_t^e &= \frac{e^{g_t^e}}{\sum_{t=1}^k e^{g_t^e}} \\ r^e &= \sum_{t=1}^k w_t^e x^{c_t^s} \end{aligned} \tag{3}$$

where $W_2 \in \mathbb{R}^{1\times(d_x^l+\sum_{i\in\{v,a,va\}} d_h^i)}$, $b_2 \in \mathbb{R}^1$ are the parameters of the multimodal sentiment word attention network.

In addition, there may not be suitable words in the candidate words. Hence, we incorporate the embedding of the special word [MASK], x^{mask}, to let the BERT model handle this problem based on the context. We then design an aggregation network to balance the contributions of the special word embedding x^{mask} and the sentiment embedding r^e. Finally, we add the r^{add} to the filtered word embedding u^l_s and obtain the refined word embedding r_l for the target word.

$$\begin{aligned} g^{mask} &= Sigmoid(W_3([r^e; x^{mask}]) + b_3) \\ r^{add} &= g^{mask} r^e + (1 - g^{mask}) x^{mask} \\ r^l &= (g^v p) r^{add} + r^v \end{aligned} \tag{4}$$

where $W_3 \in \mathbb{R}^{1 \times 2d^l_x}$, $b_3 \in \mathbb{R}^1$ are the trainable parameters.

3.3 Multimodal Feature Fusion

We describe our multimodal feature fusion module in the section and it is noted that our proposed refinement approach only modifies the textual input token embeddings, which makes it easy to be adapted for other multimodal feature fusion models based on BERT, such as MISA (Hazarika et al., 2020).

We first use the BERT model to encode the refined word embeddings $\mathbf{z}^{\mathbf{l}} = \{x^l_1, x^l_2, ..., r^l, .., x^l_{n_l}\}$ and take the representation of [CLS] as the textual representation, which is denoted as v^l. And then we use two LSTM networks to encode the visual and acoustic features and take the representations of the first words as the visual representation v^v and acoustic representation v^a. Finally, we fuse them using a non-linear layer to capture the interactions between them.

$$\begin{aligned} v^l &= \text{BERT}_{\text{textual}}(\mathbf{z}^{\text{l}}) \\ v^v &= \text{LSTM}_{\text{visual}}(\mathbf{x}^{\text{v}}) \\ v^a &= \text{LSTM}_{\text{acoustic}}(\mathbf{x}^{\text{a}}) \\ v^f &= Relu(W_4([v^l; v^v; v^a]) + b_4) \end{aligned} \tag{5}$$

where $W_4 \in \mathbb{R}^{d^f_v \times (d^l_v + d^a_v + d^v_v)}$, $b_4 \in \mathbb{R}^{d^f_v}$ are the trainable parameters of the fusion network.

We utilize a linear layer to predict the final sentiment regression labels.

$$p_f = W_5 v^f + b_5 \tag{6}$$

where $W_5 \in \mathbb{R}^{1 \times d^f_v}$, $b_5 \in \mathbb{R}^1$ are the trainable parameters of the prediction network.

Besides, to enhance the model to capture unimodal sentiment information, we use the Unimodal Label Generation Module (ULGM) (Yu et al., 2021) to generate pseudo unimodal sentiment labels and adopt them to train our model in a multi-task learning manner. For more details, we refer you to Yu et al. (2021).

4 Experiment

4.1 Datasets

We build three real-world datasets including MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek, on CMU-MOSI(Zadeh et al., 2016).

CMU-MOSI CMU multimodal opinion-level sentiment intensity (CMU-MOSI) consists of 93 videos collected from the YouTube website. The length of the videos varies from 2-5 mins. These videos are split into 2,199 short video clips and labeled with sentiment scores from -3 (strongly negative) to 3 (strongly positive). For multimodal features, we extract the visual features using Facet, which can extract the facial action units (Ekman et al., 1980) from each frame. The acoustic features are obtained by applying COVAREP (Degottex et al., 2014), which includes 12 Mel-frequency cepstral coefficients (MFCCs) and other low-level features.

However, the provided texts of the utterances in the MOSI dataset are manually transcribed from the corresponding videos by the expert transcribers, which is unrealistic for the real-world applications to obtain the texts in such a way. To evaluate the models in the real world, we replace the manually gold texts in the dataset with the texts output by the ASR models. We adopt a strong ASR model and two widely used commercial APIs to produce the texts. The utilized ASR model released by Ravanelli et al. (2021) is built on the transformer encoder-decoder framework and trained on the Librispeech dataset(Panayotov et al., 2015). The commercial APIs used by us are IBM[2] and iFlytek[3] speech-to-text APIs, which are wildly used by researchers and software developers. Finally, we apply the three ASR models to transcribe the videos into texts and construct three new datasets, namely MOSI-SpeechBrain, MOSI-IBM,

[2] https://www.ibm.com/cloud/watson-speech-to-text

[3] https://global.xfyun.cn/products/lfasr

Datasets	Models	Evaluation Metrics					
		Has0-Acc ↑	Has0-F1 ↑	Non0-Acc ↑	Non0-F1 ↑	MAE ↓	Corr ↑
MOSI-SpeechBrain	TFN(B)	68.98	68.95	69.51	69.57	115.55	48.54
	LMF(B)	68.86	68.88	69.36	69.48	117.42	48.66
	MulT(B)	71.78	71.70	72.74	72.75	109.00	54.69
	MISA	73.79	73.85	74.51	74.66	98.52	65.37
	Self-MM	73.67	73.72	74.85	74.98	90.95	67.23
	Ours	**74.58**	**74.62**	**75.70**	**75.82**	**90.56**	**67.47**
MOSI-IBM	TFN(B)	71.81	71.78	72.13	73.21	109.42	58.19
	LMF(B)	73.06	73.09	74.30	74.41	104.70	59.07
	MulT(B)	75.57	75.54	76.74	76.79	100.32	64.34
	MISA	76.97	76.99	78.08	78.17	91.23	71.30
	Self-MM	77.32	77.37	78.60	78.72	85.65	73.23
	Ours	**78.43**	**78.47**	**79.70**	**79.80**	**82.91**	**73.91**
MOSI-iFlytek	TFN(B)	71.95	72.01	72.62	72.76	107.01	56.52
	LMF(B)	71.98	72.03	72.35	72.49	106.63	59.48
	MulT(B)	77.32	77.05	78.75	78.56	89.84	68.14
	MISA	79.59	79.62	79.82	79.91	85.63	74.53
	Self-MM	80.26	80.26	81.16	81.20	78.79	75.83
	Ours	**80.47**	**80.47**	**81.28**	**81.34**	**78.39**	**75.97**
MOSI-Gold	Self-MM	82.54	82.51	84.02	84.05	72.49	78.90

Table 1: Results on the MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek datasets. (B) means the textual features are based on BERT. The best results are in **bold**.

and MOSI-iFlytek. We report the WER results of the adopted ASR models on MOSI in Appendix A. Noted that, we do not adopt MOSEI (Bagher Zadeh et al., 2018), because it does not provide the original video clips for the extracted features and annotated sentiment labels, and we can not process the original audios.

4.2 Training Details

We use Adam as the optimizer and the learning rate is 5e-5. The batch size is 64. The sentiment threshold is set to 0.5 while detecting the sentiment word position. The number of the candidate words k is 50. The other hyper-parameters of the model are reported in Appendix B. All experiments are run on an Nvidia Tesla P100 GPU. We run five times and report the average performance. The random seeds we used are 1111,1112, 1113, 1114, and 1115.

4.3 Evaluation Metrics

For the MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek datasets, following previous work (Yu et al., 2021), we take 2-class accuracy(Acc-2), F1 score(F1), mean absolute error (MAE), and correlation(Corr) as our evaluation metrics. And for Acc-2 and F1-Score, we calculate them in two ways, negative/non-negative (Non0-Acc, Non0-F1) and negative/positive (Has0-Acc, Has0-F1). As the prediction results are real values, we obtain the sentiment classification labels by mapping the sentiment scores into labels.

4.4 Baselines

We compare our proposed model with the following baselines [4]. **TFN** (Zadeh et al., 2017) uses the three-fold Cartesian product to capture unimodal, bimodal, and trimodal interactions. **LMF** (Liu et al., 2018) uses the low-rank tensors to accelerate the multimodal feature fusion process. **MulT** (Tsai et al., 2019) uses the cross-modal transformers to fuse multimodal features. **MISA** (Hazarika et al., 2020) adopts multi-task learning to map different modal features into a shared subspace. **Self-MM** (Yu et al., 2021) first generates the pseudo unimodal sentiment labels and then adopts them to train the model in a multi-task learning manner.

5 Results and Analysis

5.1 Quantitative Results

In Table 1, we show the results on the MOSI-SpeechBrain, MOSI-IBM, MOSI-iFlytek datasets.

[4] Because applying the force-alignment using the errorous ASR texts leads to cascading errors resulting in poor aligned features, we only take the models using unaligned features as our baselines for a fair comparison.

Models	Has0-Acc ↑	Non0-Acc ↑	MAE ↓
SWRM	**74.58**	**75.70**	**90.56**
w/o Position	73.59	74.57	93.67
w/o Attention	74.17	75.42	91.53
w/o Multi-modal	73.82	75.09	91.22

Table 2: Ablation analysis of our proposed model evaluated on the MOSI-SpeechBrain dataset. The best results are in **bold**.

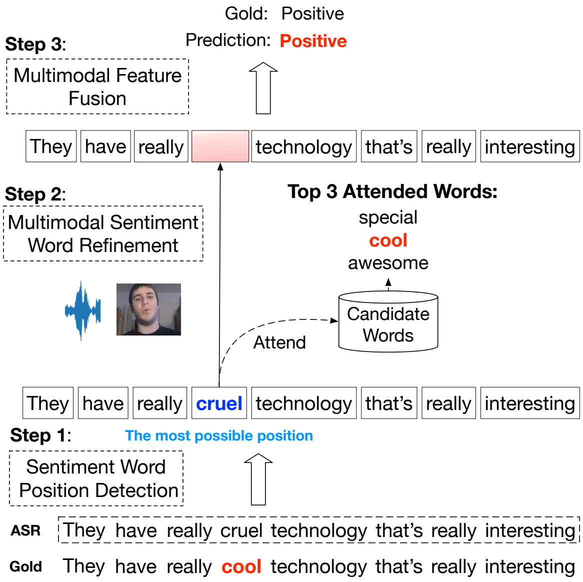

Gold: Positive
Prediction: Positive
Step 3:
Multimodal Feature Fusion
They
have
really
technology
that's
really
interesting
Step 2:
Multimodal Sentiment Word Refinement
Top 3 Attended Words:
special
cool
awesome
Candidate Words
Attend
They
have
really
cruel
technology
that's
really
interesting
Step 1:
The most possible position
Sentiment Word Position Detection
ASR
They have really cruel technology that's really interesting
Gold
They have really cool technology that's really interesting

Figure 3: Case study for the SWRM.

.

And we also list the results of the SOTA model, Self-MM, on the original MOSI dataset in the last row of the table for the performance comparison between Self-MM in the ideal world and real world. As we can see from the results, Self-MM obtains the best results on the MOSI-Gold dataset than the other datasets, which demonstrates that the ASR errors hurt the MSA models. We also observe that the better ASR model can help the MSA models achieve better performance. But it should be noted that, according to the analysis in the previous section, current ASR models still can not produce satisfactory results for the MSA models in the real world.

Comparison between the feature-based models including TFN, LMF, and MulT and finetuning-based baselines such as MISA and Self-MM, we can find that finetuning-based models obtain better results. We consider that the finetuning-based models can adapt the BERT encoder to the target task and learning more informative textual representations, which also makes them benefit more as the quality of texts increases.

Comparing to the baselines especially Self-MM, our model achieves better performance in all evaluation metrics since our model can detect the substitution error of the sentiment words and then refine the word embeddings to reconstruct the sentiment semantics in the textual modality by filtering out useless information from the input words and incorporating useful information from the candidate words generated by the language model. We also observe that the improvement of our model compared with Self-MM on MOSI-iFlytek is smaller. We consider that the main reason is fewer sentiment word substitution errors on MOSI-iFlytek.

5.2 Ablation Study

We conduct the ablation experiments to distinguish the contribution of each part. There are several different variants of our model. **SWRM** is our proposed full model. **SWRM w/o Position** does not use the sentiment word position location module and only uses the information of the special word [MASK] to dynamically refine all words. **SWRM w/o Attention** only incorporates the information of the special word [MASK] to refine the word in the multimodal sentiment word refinement module. **SWRM w/o Multi-modal** only performs the multimodal sentiment word attention and multimodal gating network based on the textual features without the acoustic and visual features.

Table 2 shows the results of the variants of our model. After ablating the sentiment word position location module, SWRM w/o Position obtains worse results than SWRM, which indicates that finding the right word for refinement is very important. The comparison between SWRM w/o Attention and SWRM w/o Position further demonstrates this conclusion. SWRM w/o Attention first detects the right position and then incorporates the information of the special word [MASK], which achieves better performance than SWRM w/o Position. But SWRM w/o Attention is still worse than SWRM, which shows using the attention network to incorporating extra information from the candidate words is useful for refinement. Comparing the SWRM w/o Multi-modal between SWRM, we can find that the model benefits from the visual and acoustic features. It is in line with our expectations since the sentiment information provided by the multimodal features can help the model detect the sentiment word and incorporate the sentiment-related information from the candidate words.

5.3 Case Study

To have an intuitive understanding of our proposed model, we show a case in Figure 3. We can see that our model first detects the most possible position based on the context and then finds that the input word in the position may be recognized incorrectly since there is a mismatch between the negative word “cruel" and either the smile or the excited tone. Hence our model decides to incorporate the related sentiment information from the candidate words to refine the word embedding. As shown in Figure 3, our model pays more attention to the candidate words "special", "cool", and "awesome". The word "cool" is exactly the gold word and the others have the same sentiment polarity as it. Beneficial from the attended candidate words, our model refines the input word and reconstructs its sentiment semantics. Finally, the refined word embeddings are fed into the multimodal feature fusion module to predict the sentiment label.

6 Conclusion

In this paper, we observe an obvious performance drop when the SOTA MSA model is deployed in the real world, and through in-depth analysis, we find that the sentiment word substitution error is a very important factor causing it. To address it, we propose the sentiment word aware multimodal refinement model, which can dynamically refine the word embeddings and reconstruct the corrupted sentiment semantics by incorporating the multimodal sentiment information. We evaluate our model on MOSI-SpeechBrain, MOSI-IBM, and MOSI-iFlytek and the results demonstrate the effectiveness of our approach. For future work, we will explore leveraging the multimodal information to detect the sentiment word positions.

Acknowledgments

This work was supported by the following Grants: National Natural Science Foundation of China (No. 62176078), National Key R&D Program of China (No. 2018YFB1005103).

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI ’17, page 163–171, New York, NY, USA. Association for Computing Machinery.

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

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

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

Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, and Sunil Kumar Kopparapu. 2018. Sentiment classification on erroneous asr transcripts: a multi view learning approach. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 807–814. IEEE.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

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*, MM ’20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1569–1576, Florence, Italy. Association for Computational Linguistics.

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

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):164–172.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1359–1367.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

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*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, Online. Association for Computational Linguistics.

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

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797.

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*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.

A WER Results on MOSI

API	SpeechBrain	IBM	iFlytek
WER	37.54	29.70	23.91

Table 3: WER results on the MOSI dataset.

B Hyper-parameter Settings

Dataset	MOSI-SpeechBrain	MOSI-IBM	MOSI-iFlytek
Batch Size	64	128	64
d_x^l	768	768	768
d_h^v	16	32	16
d_h^a	32	32	32
d_h^{va}	48	64	48
d_v^l	32	64	32
d_v^a	16	32	16
d_v^v	32	16	32
d_v^f	128	64	128

Table 4: The hyper-parameters used in training for the three datasets.