Towards Exploiting Sticker for Multimodal Sentiment Analysis in Social Media: A New Dataset and Baseline

Feng Ge **Weizhao Li** **Haopeng Ren** **Yi Cai** *
South China University of Technology
logosg@foxmail.com
se_weizhao.li@mail.scut.edu.cn
renhp_scut@foxmail.com
ycai@scut.edu.cn

Abstract

Sentiment analysis in social media is challenging since posts are short of context. As a popular way to express emotion on social media, stickers related to these posts can supplement missing sentiments and help identify sentiments precisely. However, research about stickers has not been investigated further. To this end, we present a Chinese sticker-based multimodal dataset for the sentiment analysis task (CSMSA). Compared with previous real-world photo-based multimodal datasets, the CSMSA dataset focuses on stickers, conveying more vivid and moving emotions. The sticker-based multimodal sentiment analysis task is challenging in three aspects: inherent multimodality of stickers, significant inter-series variations between stickers, and complex multimodal sentiment fusion. We propose SAMSAM to address the above three challenges. Our model introduces a flexible masked self-attention mechanism to allow the dynamic interaction between post texts and stickers. The experimental results indicate that our model performs best compared with other models. More researches need to be devoted to this field. The dataset is publicly available at https://github.com/Logos23333/CSMSA.

1 Introduction

In recent years, social media has become more and more popular, and people tend to express opinions and emotions on social media (Greenwood et al., 2016). Collecting opinions and analyzing sentiment in social media can help a lot for marketing (Alalwan et al., 2017), campaign trends prediction, and so forth. The sentiment analysis is challenging in social media because of the lack of context (Khan and Fu, 2021). Stickers attached to posts can supplement missing sentiment and help to identify sentiment precisely. Such task of performing sentiment analysis with multiple data sources is called multimodal sentiment analysis.

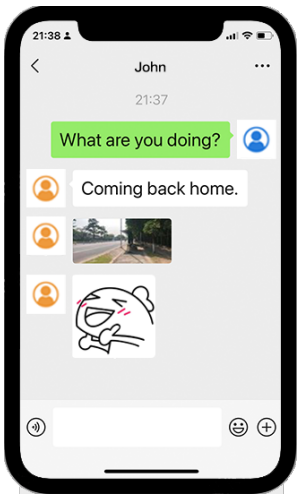

Figure 1: An example of online conversation. The first image is a photo that does not reflect any sentiment. The second is a sticker, expressing the emotional tendency of happiness. The sticker here contributes to supplementing the missing sentiment of the text.

Many research efforts have been devoted to multimodal sentiment analysis (Poria et al., 2015). However, most of the multimodal sentiment analysis datasets are based on real-world photos (Yu et al., 2020). Research focusing on sticker-based multimodal sentiment is limited. In social media, people tend to express the sentiment with stickers rather than real-world photo (Wang et al., 2019). Stickers can convey more vivid and direct emotion. For example, as shown in Figure 1, the first image is a photo related to the post context, but it does not reflect any sentiment. Instead, the second image, i.e., the sticker, shows a sense of happiness, revealing the missing sentiment of the corresponding texts. To this end, we introduce CSMSA, a challenging **C**hinese **S**ticker based **M**ultimodal dataset for **S**entiment **A**nalysis task in social media. CSMSA includes 28k text-sticker pairs, including 1.5k sentiment-annotated pairs, and 16k different stickers. This dataset is the first annotated sticker-based dataset for multimodal sentiment analysis to

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 6795–6804
October 12–17, 2022.

the best of our knowledge. Our dataset will release to encourage research on multimodal sentiment analysis in social media.

The proposed CSMSA task is challenging in three aspects: First, stickers may be inherently multimodal because they are embedded with texts[1], while other datasets have only real-world photos. For example, as shown in Figure 2, sticker (a) is attached with the text "I'm beat" and sticker (b) is attached with the text "I'm so touched". The difficulty is that the same sticker with different sticker texts may vary significantly in sentiment. Second, stickers are highly different in styles, leading models to learn robust representations for the stickers following various distributions hardly (Huo et al., 2018). In contrast, most of the photo-based datasets only consist of human faces or food (Hasan et al., 2019), and the styles do not differ much. Considering the large inter-series variation of stickers, the CSMSA task requires models to adapt to different artistic styles. Third, the sentiment fusion of text and stickers is complex. For example, the text sentiment in Figure 2 is negative obviously. However, after combining the text and the sticker (b), the multimodal sentiment is positive, showing a sense of moving and touching. Multimodal sentiment does not always favor a single modality, making the fusion between modalities complex. In general, the first and second challenges are due to the difficulty of modeling the stickers, making it more challenging to fuse the sentiment with the post text.

We propose a Sticker-Aware Multimodal Sentiment Analysis Model (SAMSAM) to address the above three challenges. To address the first and second challenges, SAMSAM introduces the sticker text and sticker series tag, which can be seen as a further complement to sticker sentiment to model the sticker wholly and accurately. To address the third challenge, SAMSAM adopts the flexible masked attention mechanism to allow post texts and stickers to interact fully but selectively. The masked attention mechanism can help the model extract the most helpful information for the current sentiment judgment. We conduct experiments on the proposed CSMSA dataset. The experimental results show that our model performs best in the challenging CSMSA task. The ablation study indicates that the encoding of sticker texts and sticker series help in understanding multimodal sentiment,

[1]In the following, we call the text embedded in the sticker "sticker text" to avoid confusion with the post text.

Figure 2: An example for showing how the sticker influences sentiment in multimodal sentiment analysis. A sentence accompanied by different stickers may express reversed sentiments. An image accompanied by different sticker texts may also express reversed sentiments. The contents in brackets are translations from Chinese.

and the proposed masked self-attention mechanism can improve model performance significantly.

The contributions of this paper are as follows:

- We analyze the difference between real-world photos and stickers. We reveal three challenges of the CSMSA task: inherent multimodality, significant inter-series variation, and complex sentiment fusion.
- We propose a sticker-based human-annotated dataset. The dataset aims to test the model's ability to leverage stickers to supplement the missing sentiments of texts. To the best of our knowledge, we are the first to focus on stickers in multimodal sentiment analysis.
- We propose SAMSAM and step towards sticker-aware multimodal models. We conduct experiments against other models. Experimental results indicate that our model performs best compared with other models.

2 Related Work

2.1 Multimodal Sentiment Analysis Datasets

Multimodal sentiment analysis has attracted more and more attention recently. Due to the diversity of modalities and interactions between modalities, researchers have proposed a wide variety of multimodal sentiment analysis datasets, adapting to different real-world scenarios. Bagher Zadeh et al. (2018) propose CMU-MOSEI, one of the enormous datasets for sentiment analysis and emotion recognition with three modalities. The data of CMU-MOSEI came from YouTube monologue videos,

and the dataset was both sentiment and emotion annotated. Castro et al. (2019) propose a multimodal sarcasm detection dataset (MUStARD), which is compiled from famous TV shows and consists of audiovisual utterances annotated with sarcasm labels. Furthermore, Hasan et al. (2019) proposes a UR-FUNNY dataset, which contains text-audio-video three modalities for understanding humor. Yu et al. (2020) propose a Chinese multimodal sentiment analysis dataset with both multimodal and independent unimodal annotations. The primary difference between CSMSA and previous studies is that traditional multimodal datasets focus on videos of a speaker with their face, and few researchers pay attention to the sticker. Our proposed CSMSA dataset focuses on the sticker, studying the sentiment interaction between text and sticker.

2.2 Multimodal Sentiment Analysis Methods

Multimodal sentiment analysis mainly focuses on utilizing multiple resources to predict human emotions. Most multimodal models focus on three modalities: acoustic, visual, and text. Han et al. (2021b) propose Multimodal-infomax model. They propose a two-level mutual information maximization and design an entropy estimation module to facilitate the computation of Barber-Agakov (Agakov, 2004) lower bound and the training process. Hazarika et al. (2020) propose a novel framework MISA, which projects each modality to two distinct subspaces, reducing the modality gap and capturing characteristic features. Han et al. (2021a) propose Bi-Bimodal Fusion Network, which learns two text-related pairs of representations and generates the final prediction through the concatenation of four head representations.

2.3 Sticker-Related Research

Gao et al. (2020) propose to recommend an appropriate sticker to the user based on multi-turn dialog context history without any external labels. They release a dataset of 340K multi-turn dialog and sticker pairs. The dataset contains the most significant number of stickers available for sticker recommendation tasks. Fei et al. (2021) propose a new task Meme incorporated Open-domain Dialogue (MOD). The memes mentioned in (Fei et al., 2021) can be seen as the same thing with the sticker. A large-scale open-domain multimodal dialogue dataset incorporating great Internet memes into utterances is also released with the sentiment label of memes. However, their dataset is constructed by inserting memes as appropriately as possible into the multi-turn conversation dataset. This construction does not reflect the real interaction and supplementary between modalities because the memes inserted must not be too much different from the sentiment of the text. In the real world, the sentiment of stickers may be very different from text, and the sentiment fusion of text and sticker is very complex. For example, in the second example of Figure 3, the sentiment of the sticker is the opposite of the text, indicating that the sticker is not just complementary and it can play a leading role in multimodal sentiment analysis. In general, our dataset is more consistent with the data distribution of the real world.

3 Dataset

3.1 Data Collection

We collect the dataset from one of the most popular messaging apps. A huge number of stickers are released in this app, and everyone can use these stickers easily and conveniently in private or group chat. In this app, a sticker has a different identifier than other typical images, which allows us to focus on stickers and filter noises (e.g., screenshots and photos). All stickers are resized to a uniform size of 224 x 224 pixels.

We select eight public open chat groups consisting of active members, and the chat history of these groups is collected. To get multimodal data containing both text and sticker, we traverse every sticker in the chat history and collect the context of the sticker. The sticker and the text must be sent by the same person. Due to privacy concerns, we also filter out user information and anonymize user IDs.

3.2 Annotation

We employ four well-educated annotators to label the sentiments of text-sticker pairs. The annotators are asked to judge the text sentiment, image sentiment, multimodal sentiment, and *image_can_help* label. The *image_can_help* label indicates if the given sticker can assist in judging multimodal sentiment. This label can help researchers investigate how stickers' modality supplements the sentiment of the text. A dataset with multimodal and independent unimodal annotations allows researchers to study the interaction between modalities (Yu et al., 2020). Thus, our dataset can fully support unimodal sentiment analysis research.

To study the differences across sticker series, we

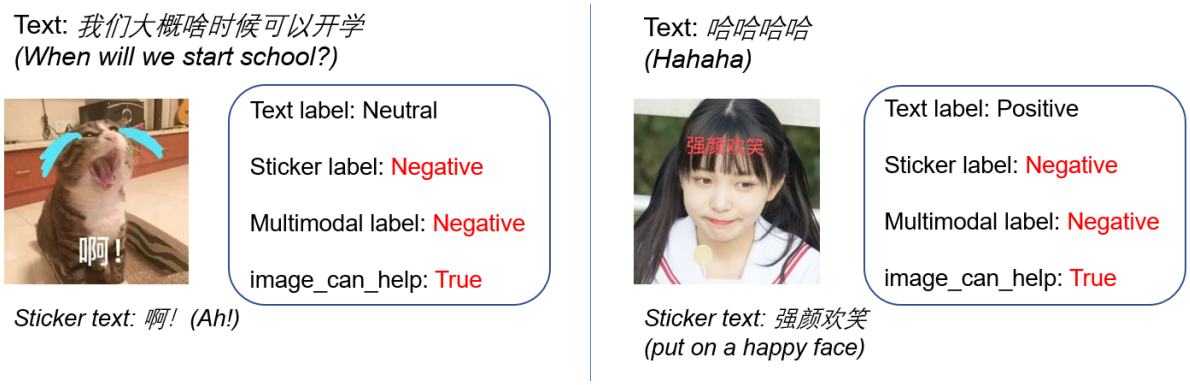

Figure 3: Two examples of the CSMSA dataset illustrate the sticker's effectiveness in sentiment analysis. For each text-sticker pair, in addition to multimodal annotations, the CSMSA dataset has independent unimodal annotations and an *image_can_help* label. The *image_can_help* label indicates whether the sticker can help to analyze sentiment.

Datasets	Size	# Sti.	# Ano.
MOD	45k	307	0
StickerChat	**340k**	174k	0
CSMSA	28k	16k	1.5k

Table 1: Comparison with other sticker-based datasets. # Sti. represents the number of stickers. # Ano. represents the number of samples with human-annotated multimodal sentiment label.

Task	Train	Valid	Test
Easy Task	942	314	314
Hard Task-1	1297	127	146
Hard Task-2	1290	130	150
Hard Task-3	1373	109	88

Table 2: The split statistics of the CSMSA dataset.

asked the annotators to group stickers into difference series. These sticker series include stickers with a similar style. Each annotator will give a confidence score used to calculate a final sentiment score considering the different views of all the annotators. The original data are randomly assigned to each annotator. Four annotators will decide every instance for the sake of the quality of the labeling. We weigh the confidence score of each annotator for the same label, and the label with the highest score will be preserved.

3.3 Statistics and Analysis

The final dataset contains 28k text-sticker pairs and 16k different stickers. The statistical comparison of the CSMSA dataset with existing sticker-based datasets is shown in Table 1. This dataset is the first annotated sticker-based dataset for multimodal sentiment analysis to the best of our knowledge. Our study is conducted on 1.5k annotated data, and the large-scale dataset is for future work.

We also conduct studies on the ability of text and images to express sentiment by using annotated unimodal labels. According to our statistics, nearly 52.1% of texts in the CSMSA dataset do not convey any sentiment, while sticker has a 73% probability of conveying sentiment. Meanwhile, 591 (37.6%) of the data are marked as *image_can_help* in the whole dataset. It means that if the stickers modality is ignored, about 37.6% of the data may not accurately determine their sentiment.

3.4 Case Analysis

Stickers can help to analyze sentiment. However, previous work considers only real-world photos, ignoring stickers heavily used in social media. In Figure 3, we show two examples of the CSMSA dataset to reveal how stickers can assist in determining sentiment. In the first example, the post text asks about the time of the school year without expressing any sentiment, while the sticker's sentiment is negative. The sticker creator attaches text "Ah!" to the sticker and adds blue "tears" to the cat, giving it a strong emotional impact. In the second example, the sentiment of the text is positive. However, we can know that it is a forced smile with the help of the sticker, and the multimodal sentiment is negative, which is the opposite of the text label.

3.5 Dataset Division

There are different series of stickers, and each series has an individual artistic style. Compared to the limited number of emojis, the number of stick-

Figure 4: An example shows the division of Hard Task-3 and the significant variation of styles between the stickers series.

ers is huge, and new stickers appear every day. In order to investigate whether the model can adapt to the newcomer stickers and the artistic style among different sticker series, we divide the dataset into four different ways: Easy Task, Hard Task-1, Hard Task-2, and Hard Task-3. The division is shown in Table 2. In particular, Easy Task randomly divides the whole dataset in the ratio of 6:2:2. The training set and validation set stickers may appear in the test set. Hard Task-1 randomly divides the stickers so that stickers from the test set will not appear in the training and validation sets. Hard Task-2 randomly divides the sticker series so that the sticker series from the test set will not appear in the training and validation sets. Hard Task-3 partitions the stickers into three categories: human face-based, pet-based, and cartoon character-based. As shown in Figure 4, there is a considerable difference in style between them. The Hard Task-1 measures whether the model can adapt to the new sticker. The Hard Task-2 measures whether the model can adapt to the new sticker series. The Hard Task-3 measures whether the model can adapt to unseen sticker series, which varies greatly from the previous one.

4 Model

4.1 Task Definition

Formally, given an post text $X = (x_1, x_2, \ldots, x_m)$, a sticker I, the sticker text $S = (s_1, s_2, \ldots, s_n)$ and the sticker series E, the CSMSA task requires model to predict the multimodal sentiment label $y \in \{Positive, Negative, Neutral\}$. x_i denotes the i-th word in the post text, s_i denotes the i-th word in the sticker text, m is the length of post text, n is the length of sticker text.

4.2 Feature Layer

Post Text Encoder. BERT (Devlin et al., 2019; Sun et al., 2019) is a transformer-based pre-trained model which uses a Masked Language Model to predict randomly masked or replaced words. We use the bert-base-Chinese[2] weights provided by google to fine-tune the proposed model. For classification tasks, BERT takes the final hidden state of the first token [CLS], i.e., h, as the representation of the whole sequence.

$$h = BERT(X) \tag{1}$$

Sticker Encoder. ResNet (He et al., 2015) is a classic neural network used as a backbone for many computer vision tasks. We use the ResNet34 as our image encoder to obtain representations of stickers. An input sticker I is resized to 224×224 and then sent through the ResNet model to obtain sticker representations f.

$$f = ResNet(I) \tag{2}$$

Sticker Text Encoder. The text within a sticker is an essential component for understanding the sticker. To extract the textual information contained in sticker, we introduce PaddleOCR (Du et al., 2020) to recognize texts S within a sticker. The text feature u of sticker text S is obtained through LSTM (Hochreiter and Schmidhuber, 1997), and this text feature can assist in fusing multimodal information and determining the final sentiment.

$$u = LSTM(S) \tag{3}$$

Sticker Series Embedding. The differences in artistic styles cause significant inter-series variations, making the model difficult to understand stickers. For the model to distinguish the different artistic styles of the series explicitly, the sticker series tag e is fed as a embedding.

$$v = Embedding(E) \tag{4}$$

4.3 Interaction Layer

There are four types of information in the interaction layer: post text h, sticker f, sticker text u, and sticker series v, which make up a tuple (h, f, u, v). Accordingly, four inter-information relations need to be considered:

[2] https://huggingface.co/bert-base-chinese

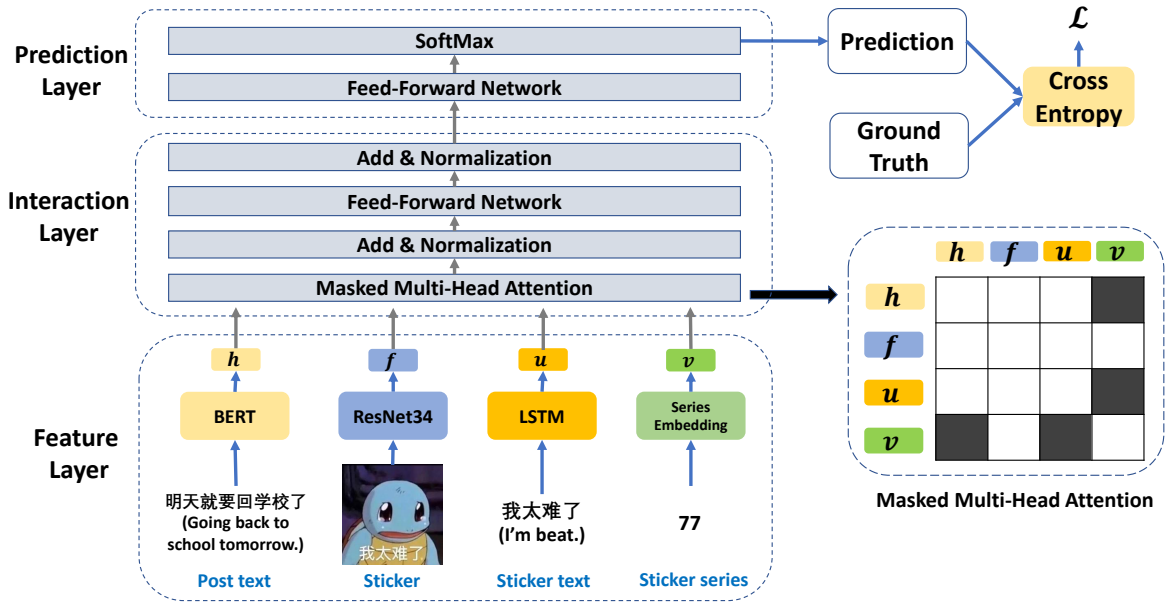

Figure 5: The overview of the proposed SAMSAM and masked multi-head attention. SAMSAM consists of three parts, the feature layer, the interaction layer, and the prediction layer.

(1) The alignment between h and f is beneficial to fusing the sentiment between post text and sticker, we need to find the most relevant text-image information according to the relations between h and f.

(2) As the sticker text is an essential way of expressing emotion, we need to find the most relevant sticker information according to the relations between h and u.

(3) Similar to (1) and (2), we want to fuse the sentiment of the sticker and find the most relevant sticker information according to the relations between f and u.

(4) We also want to let the model understand the current sticker series style according to the relations between f and v.

We adopt multi-head attention in the interaction layer considering the above factors. Multi-head attention allows $h-f$, $h-u$, $f-u$ and $f-v$ pair to interact fully. Since post text and sticker text are not intrinsically related to sticker series, we need to prevent some of the information from interacting explicitly for noise mitigation. Thus we adopt a well-designed attention mask $\mathcal{M}$ in addition to the original multi-head attention. The overview of the interaction layer and mask is shown in Figure 5.

In the interaction layer, we employ three feed-forward networks with different parameters to project the C_i into three different spaces:

$$Q_i = FN(C_i),\ K_i = FN(C_i),\ V_i = FN(C_i) \tag{5}$$

C_i represents the i-th source of C, i.e, h, f, u and v. The model then takes each Q_i to attend to K_j, and uses the attention weights $\alpha_{i,j}$ to gain the weighted sum of V_i:

$$C'_i = \sum_{j=1}^{4} \alpha_{i,j} * V_i \tag{6}$$

$$\alpha_{i,j} = \frac{exp(Q_i * K_j \odot \mathcal{M})}{\sum_{k=1}^{4} exp(Q_i * K_k \odot \mathcal{M})} \tag{7}$$

4.4 Prediction Layer

To aggregates information in different sources, we stack the hidden states in C' as the hybrid output: $O = [C'_1; C'_2; C'_3; C'_4]$. Finally, we feed O to a one-layer feed-forward network followed by a softmax function for the sentiment prediction distribution:

$$P(y|X, I, S, E) = softmax(FFN(O)) \tag{8}$$

To optimize all the parameters in our SAMSAM, the objective is to minimize the standard cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} logP(y^{(c)}|X^{(c)}, I^{(c)}, S^{(c)}, E^{(c)}) \tag{9}$$

where $\mathcal{D}$ denotes all samples in the dataset.

5 Experimental Setup

5.1 Compared Methods

BERT BERT is one of the state-of-the-art methods to address classification tasks. We implement BERT as a text-only baseline.

BERT-ST The input sentence of BERT-ST is the concatenation of post text and sticker text, separated by [SEP]. We implement BERT-ST as a text-only baseline.

RoBERTa RoBERTa is an optimized method for BERT. We implement RoBERTa as a strong text-only baseline.

ResNet34 We implement ResNet34 as a image-only baseline model.

mBERT Yu and Jiang (2019) design mBERT on top of the baseline BERT architecture. They directly concatenate the image features with the final hidden states of the post text.

MMTF Li (2021) propose a Multimodal Transformers (MMTF) for Meme Classification task. They propose a multimodal attention layer to fully interact with the text and image based on the cross-attention mechanism. We adopt MMTF as a multimodal baseline.

5.2 Implementation Details

All the models are trained in 100 epochs with an NVIDIA GTX 2080Ti, and we use the best model on the validation set for evaluation. We use Pytorch (Paszke et al., 2017) and HuggingFace's transformers (Wolf et al., 2020) to implement our model. AdamW (Loshchilov and Hutter, 2019) optimizer is used to optimize our model with learning rate 5e-7. The batch size is 6.

6 Experimental Results

6.1 Overall Performance

Table 3 shows the experimental results of our comparison with other models and ablation study.

The experimental results show that SAMSAM achieves the best results among multimodal and unimodal models. Multimodal models MMTF and mBERT are not well adapted to the CSMSA task because they ignore the sticker text embedded in the sticker, and the styles of stickers vary significantly. Unimodal models perform poorly because they focus on a single modality and ignore the complementary effect between modalities.

The text-only BERT and RoBERTa models do not consider the sentiment of the stickers, resulting in their poor performance. The results of BERT-ST are worse than BERT because of two reasons. The first reason is the error propagation caused by the inaccurate recognition of OCR. The second is that the sentiment of the sticker must be combined with the image feature. Otherwise, the wrong sticker sentiment will further affect the multimodal sentiment judgment. The image-only model ResNet34 does not consider the text's sentiment, so it performs worse than the other multimodal models. The image-only models perform better than the text-only models on Easy Task and Hard Task-1 because the short text lacks context, and it is not easy to predict the sentiment, while the sentiment of the sticker is often stronger and more direct than the text. The image-only models perform worse on Hard Task-2 and Hard Task-3 because models fail to adapt to the new sticker series.

Generally, we can see that multimodal and image-only models perform worst on Hard Task-3 and best on Easy Task. This result also validates the challenge of the CSMSA task. It is difficult to capture and recognize stickers' sentiments because of the large inter-series variation of stickers. Models failed to generalize well when a new sticker series appeared, making it lousy to predict multimodal sentiment. Unfortunately, new sticker series appear every day. The CSMSA task is challenging and needs to be further researched.

6.2 Ablation Study

As shown in Table 3, the accuracy and F1 score decreased on almost all tasks when the mask, post text, sticker series, and sticker text were removed, respectively. The mask helps the model extract the most helpful information and eliminate some noise. The sticker text helps the model understand the sticker's sentiment and thus correctly predict multimodal sentiment. The series embedding helps the model capture the stylistic features of each series. Similarly, post text and image are also essential for multimodal sentiment judgment.

6.3 Case Study

We show some analysis examples in Figure 6. In the first example, because the post text has the word "reported", its sentiment is Negative, which leads to the misjudgment of the text-only model BERT. We know that the sticker sentiment is Positive because of the blush and smile, and the sticker text

	Methods	Easy Task		Hard Task-1		Hard Task-2		Hard Task-3	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Text-only	BERT	0.6178	0.5764	0.5274	0.4709	0.5467	0.5102	0.5114	0.3617
	BERT-ST	0.6146	0.5609	0.4932	0.4772	0.5333	0.4769	0.5227	0.3814
	RoBERTa	0.6210	0.5941	0.5479	0.4892	0.5133	0.5216	0.5341	0.4194
Image-only	ResNet34	0.6178	0.5701	0.5753	0.5530	0.5342	0.4778	0.4886	0.3484
Multimodal	mBERT	0.5924	0.5576	0.5753	0.5156	0.5600	0.5361	0.5341	0.4037
	MMTF	0.5955	0.5374	0.5479	0.4886	0.5267	0.4876	0.5227	**0.4428**
	SAMSAM	**0.6369**	0.6180	**0.5959**	**0.5669**	**0.5533**	0.5179	**0.5455**	0.4265
Ablation Study	w/o MASK	0.6306	0.6060	0.5685	0.5483	0.5133	0.4929	0.4773	0.3710
	w/o IMG	0.6274	**0.6199**	0.5685	0.5467	0.5200	**0.5251**	0.5114	0.3913
	w/o PT	0.6210	0.5665	0.5411	0.4845	0.5267	0.4924	0.5114	0.3782
	w/o SE	0.6146	0.5594	0.5685	0.5112	0.5467	0.4978	0.5227	0.4133
	w/o ST	0.6242	0.6179	0.5890	0.5664	0.5267	0.5006	0.5117	0.3925

Table 3: Overview of the experimental results. Acc. represents accuracy, and the F1 represents the weighted F1 score. MASK represents the mask mechanism. IMG represents the image feature. PT represents the post text. SE represents the sticker series embedding, and ST represents sticker text.

Post Text: *我直接举报了*
(I reported it directly.)

Sticker text: 让人怪不好意思的
(Sorry not sorry.)

× BERT: Negative
√ ResNet: Positive
× MMTF: Negative
√ SAMSAM: Positive
Ground Truth: Positive

Post Text: *你要做什么呀*
(What do you want to do?)

Sticker text: *一个耿直的微笑*
(A straight smile.)

× BERT: Neutral
× ResNet: Neutral
× MMTF: Neutral
√ SAMSAM: Positive
Ground Truth: Positive

Figure 6: Examples of multimodal sentiment analysis produced by different models on CSMSA dataset.

"Sorry and not sorry" here also further verifies the sentiment. After combining the text and sticker, we know that although the text says about "report" people and conveys some negative emotions. However, the user is happy with the situation rather than sad. In the second example, the post text does not reflect any emotion, indicating that the sentiment label is Neutral. The expression of the dog in the sticker is very subtle, in which there has a smile but not obvious. After combining the word "smile" in the sticker text, we can learn that the sticker's sentiment is Positive. The text-only model BERT fails to consider the sticker sentiment, resulting in the wrong judgment. The image-only model and MMTF are biased in judging the sticker's sentiment because they failed to combine the sticker text, which led to their incorrect judgment of multimodal sentiment. After combining the sticker text, our model correctly predicts the multimodal sentiment, i.e., Positive.

7 Conclusion

In this paper, we propose to focus on stickers rather than the real-world photo in the field of multimodal sentiment analysis. The sentiment fusion of stickers and texts is complex and challenging. Compared with real-world photos, the sticker is inherently multimodal, and it has a significant inter-series variation, making it difficult to encode. We propose a sticker-based dataset for the sentiment analysis task, with 1.5k text-sticker human-annotated pairs. This dataset is the first annotated sticker-based dataset for multimodal sentiment analysis to the best of our knowledge. Experimental results validate the challenges of the CSMSA task. Previous models cannot be directly applied to this task. More researches need to be devoted to this field.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62076100), and Fundamental Research Funds for the Central Universities, SCUT(x2rjD2220050), the Science and Technology Planning Project of Guangdong Province (2020B0101100002).

References

David Barber Felix Agakov. 2004. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201.

Ali Abdallah Alalwan, Nripendra P. Rana, Yogesh K. Dwivedi, and Raed Algharabat. 2017. Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics*, 34(7):1177–1190.

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 *ACL*, pages 2236–2246.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *ACL*, pages 4619–4629.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. PP-OCR: A practical ultra lightweight OCR system. *CoRR*, abs/2009.09941.

Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *CoRR*, abs/2109.01839.

Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. 2020. *Learning to Respond with Stickers: A Framework of Unifying Multi-Modality in Multi-Turn Dialog*, page 1138–1148. Association for Computing Machinery.

Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. Social media update 2016. *Pew Research Center*, 11(2):1–18.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. *Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis*, page 6–15. Association for Computing Machinery.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *EMNLP-IJCNLP*, pages 2046–2056.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. *MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis*, page 1122–1131. Association for Computing Machinery.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2018. Webcaricature: a benchmark for caricature recognition. In *British Machine Vision Conference*.

Zaid Khan and Yun Fu. 2021. *Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation*, page 3034–3042. Association for Computing Machinery, New York, NY, USA.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multimodal transformers for meme classification in Tamil language. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 352–356.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, pages 2539–2544.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206. Springer International Publishing.

Yuan Wang, Yukun Li, Xinning Gui, Yubo Kou, and Fenglian Liu. 2019. Culturally-embedded visual literacy: A study of impression management via emoticon, emoji, sticker, and meme on social media in china. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org.

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*, pages 3718–3727.