Impact of Stickers on Multimodal Chat Sentiment Analysis and Intent Recognition: A New Task, Dataset and Baseline

Anonymous ACL submission

Abstract

Stickers are increasingly used in social media to express sentiment and intent. When finding typing troublesome, people often use a sticker instead. Despite the significant impact of stickers on sentiment analysis and intent recognition, little research has been conducted. To address this gap, we propose a new task: **M**ultimodal chat **S**entiment **A**nalysis and **I**ntent **R**ecognition involving **S**tickers (MSAIRS). Additionally, we introduce a novel multimodal dataset containing Chinese chat records and stickers excerpted from several mainstream social media platforms. Our dataset includes paired data with the same text but different stickers, and various stickers consisting of the same images with different texts, allowing us to better understand the impact of stickers on chat sentiment and intent. We also propose an effective multimodal joint model, MMSAIR, for our task, which is validated on our datasets and indicates that visual information of stickers counts. Our dataset and code will be publicly available.

1 Introduction

With the popularization of social media, increasing number of users have turned it into significant mediums to express their sentimental trends (Gaind et al., 2019) and behavioral intents (Purohit et al., 2015). Sentiment analysis aims to determine whether the user is positive, negative, or neutral. Intent recognition, on the other hand, focuses on identifying the intent category. Both of these tasks are crucial in the field of Natural Language Understanding (NLU). Most of the time, sentiment and intent occur simultaneously, with sentiment driving the generation of some intents, and intents in turn revealing a certain sentiment (Lewis et al., 2005).

In chatting applications, social platforms, and media comment sections, a plethora of images, commonly referred to as stickers, emoticons, emojis or memes[1], can be observed. These images serve as a substitute for expressing thoughts that are challenging to convey by text alone, aiding individuals in better expressing sentiment and intent (Ge et al., 2022). However, this field hasn't been extensively researched due to issues such as text-image misalignment and lack of suitable datasets.

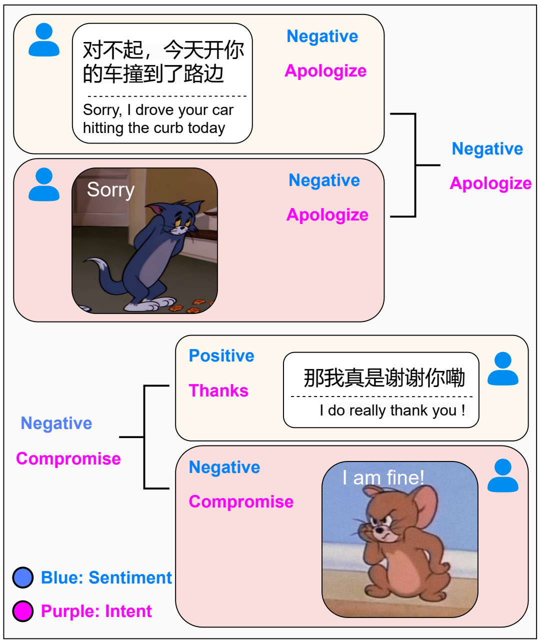

Figure 1: A chat record from a social media platform. Only by combining stickers can we discern the true pessimism and complaint the second man wants to express.

Currently, numerous studies have separately investigated multimodal sentiment analysis (Abdullah et al., 2021) and intent recognition (Huang et al., 2023). However, a handful have combined these two tasks. Most studies explore the fusion of modalities like real photos and text (Yang et al., 2019), video and text (Seo et al., 2022), audio and video with text (Akbari et al., 2021), etc., with limited research focusing on stickers and chat text. In social media, people prefer using stickers to express themselves. Stickers are often more convenient compared to text, allowing for a vivid and

[1]In this paper, we collectively refer to them as "stickers".

direct expression of ideas (Rong et al., 2022). As shown in Figure 1[2], the text and sticker send by the first man both convey a negative sentiment and an intent to apologize, making it easy to comprehend his overall message. In contrast, while the text send by the second man indicates optimism and gratitude, the sticker shows a sense of pessimism and resignation, implying that he desires to express negativity and complaints. In such situations, when it might be difficult to express directly through language, a sticker can easily convey inner feelings. Thus, only by considering stickers simultaneously can we accurately determine sentiment and intent. In addition, it can be seen that sentiment and intent are interrelated. Although the context seems to express gratitude, it is clear that the intent cannot be gratitude after receiving a negative sentiment, so it must be a compromise. Similarly, based on the intent of compromise, it can be seen that the sentiment is definitely negative. Therefore, sentiment and intent need to be handled together. Consequently, we introduce Multimodal chat Sentiment Analysis and Intent Recognition involving Stickers (MSAIRS), a completely new task, as well as a dataset of the same name to support our research.

MSAIRS task is challenging because pairing the same text with different stickers can yield different outcomes. Moreover, stickers are often multimodal themselves, containing both image and text[3], leading to variations when the image remains the same but sticker-text differs. Therefore, the task requires adept handling of context, stickers, and sticker-texts, demanding valid multimodal fusion methods (Zhang et al., 2021). To address these challenges, we introduce a simple and effective baseline: a joint Model for Multimodal Sentiment Analysis and Intent Recognition (MMSAIR). MMSAIR separately processes the input context, sticker and sticker-text, and utilize multi-head masked attention mechanisms to integrate these components, ultimately enabling combined sentiment analysis and intent recognition. Experimental results show that compared with many unimodal and multimodal models, our multimodal model performs better, indicating the necessity of simultaneously integrating the visual information of stickers.

The contributions of this paper are as follows:

- We introduce MSAIRS task and dataset to investigate the impact of stickers on multimodal chat sentiment and intent.

- We introduce a novel multimodal baseline, MMSAIR, for the joint task of multimodal sentiment analysis and intent recognition.

- To the best of our knowledge, we are the first to investigate the joint task of multimodal sentiment analysis and intent recognition involving stickers.

[2]Corresponding English translation is below the dotted line, as is the case with the other images in this paper.

[3]For differentiation, we refer to the chat text as "context", and the text within the stickers as "sticker-text".

2 Related Work

The emergence and rapid dissemination of stickers on the internet has given rise to numerous related studies(Shifman, 2013; Tang and Hew, 2019). From a sociological perspective, stickers are defined as valuable cultural units and symbols, representing a distinctive feature of global life in the internet age(Iloh, 2021).

The popularity of stickers mainly lies in the fact that they often contain rich metaphorical content, which reflects the users' sentiments and intents. (French, 2017) examined the sentimental correlation between the implicit semantics of stickers and the textual content of social media discussions, while emphasizing the significance of stickers in social media sentiment analysis. (Prakash and Aloysius, 2021) employed neural networks for facial recognition in memes containing human portraits, facilitating sentiment analysis of stickers. However, not all stickers incorporate human portraits. (Pranesh and Shekhar, 2020) proposed sentiment analysis of stickers with different styles based on transfer learning. They conducted unimodal and bimodal sentiment analysis on the textual descriptions and images within stickers. To delve deeper into the intrinsic sentiments of stickers, several studies have analyzed stickers with specific emotion such as hatred and humor. (Lestari, 2019) analyzed the ironic attributes of internet stickers from a linguistic perspective. (Tanaka et al., 2022) constructed a memes humor analysis dataset containing 7500 stickers. They contend that the humor attribute of stickers originates from the incongruity between stickers and captions, and they conducted humor analysis of stickers based on this theoretical framework. (Qu et al., 2023) analyzed the implicit hateful emotions within stickers, indicating that the intentions of individuals who post hateful stickers can have adverse real-world consequences.

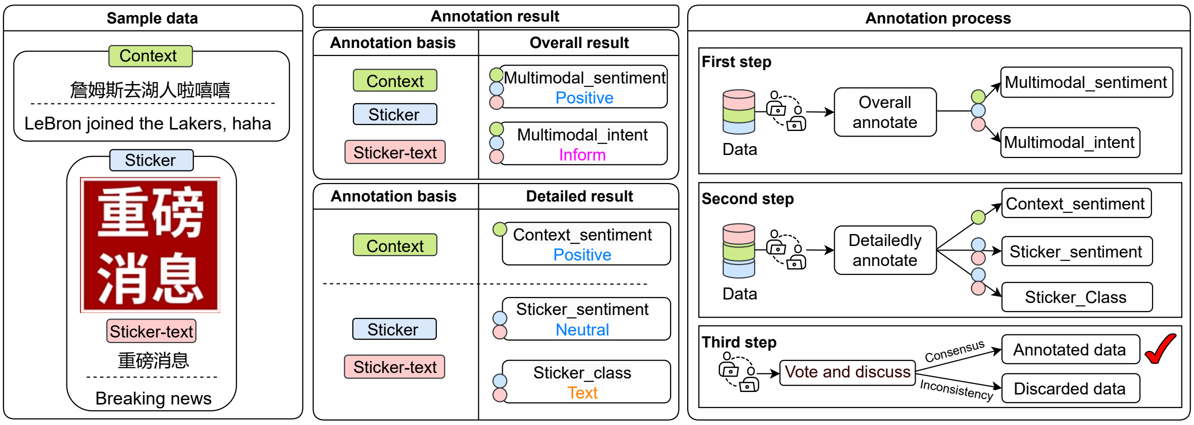

Figure 2: Annotating process and the annotation results obtained using this process for a sample from our dataset.

In fact, all stickers inherently carry a certain level of intent, and sentiments serve to amplify these intentions(Saha et al., 2021).To investigate the underlying intent behind stickers in social media, (Jia et al., 2021) introduced a dataset for recognizing intent behind social media post images. This task aims to recognize the intent behind images posted on social media by real individuals. (Xu et al., 2022) introduced a large-scale and comprehensive sticker dataset, encompassing labels such as sticker subjects, metaphors, aggressiveness, and emotions. However, current intent recognition of social media stickers solely considers unimodal information and overlooks the strong correlation between sentiment and intent. To address these gaps, we propose a multimodal sentiment analysis and intent recognition dataset tailored for social media stickers.

3 MSAIRS Dataset

3.1 Data Preparation

To study the sentiment and intent in multimodal chat conversations with stickers, we introduce the MSAIRS dataset. Referring to the CSMSA dataset (Ge et al., 2022), MSAIRS retains the sentiment labels while adding multimodal intent labels. Our research team manually collect over 5k chat records or comments with clear intent and stickers from social media platforms such as WeChat, Tik Tok and QQ. For each data entry, we ensure that both the context and sticker are sent by the same individual and apply anonymization procedures to safeguard user privacy. For stickers containing text, we utilize PaddleOCR (Du et al., 2020) to automatically extract the sticker-text and then add it to our dataset.

3.2 Data Annotation

We employ five linguistics professionals, each with rich experience in annotating Chinese datasets. The detailed annotation process and example are shown in Figure 2. Each annotator is required to label five categories. The context_sentiment and sticker_sentiment are the sentiment label separately analyzing the context and sticker, while the multimodal_sentiment is the overall sentiment analysing both the context and sticker. For intent labels, referring to Mintrec (Zhang et al., 2022), we replace several inappropriate labels with some more common labels on social media, categorizing them into twenty classes as listed in Table 1. In Figure 2, the context might be informing, flaunting or taunting. Only considering the sticker can we clearly know that someone is informing others of the news. This situation creates difficulty in determining specific intent from a single modality, so intent labels are only assigned to the multimodal_intent. The sticker_class label represents the type of sticker, broadly categorized into four classes: real person, real animal, virtual entity (e.g., cartoon), and text-only, for further study on stickers.

3.3 Manual and Automatic Review

To ensure data accuracy and credibility, each entry is retained only if three or more professionals annotate all the same labels. Otherwise, the entry is discarded. Then regarding the data where all annotators do not agree, we engage in discussions to strive for consensus. If consensus can't be reached, the data is discarded. After manual review, we retain 3.5k pieces with consistent labels.

Then we use the GPT4-Vision model to categorize the overall sentiment and intent of the text

Intent	Brief description	Quantity	Proportion
Comfort	Relax someone by making them feel at ease.	103	3.30%
Oppose	Disagree with or hinder someone or something.	173	5.55%
Greet	Meet or welcome someone with a kind wave, smile or word.	106	3.40%
Complain	Express discontent or annoyance to someone or something.	278	8.92%
Ask for help	Express the need for someone's help or guidance.	166	5.32%
Taunt	Use irony or sarcasm to mock someone.	158	5.07%
Apologize	Regret a mistake and request forgiveness.	**58**	**1.86%**
Introduce	Detailedly describe or recommend someone or something.	199	6.38%
Guess	Speculate on the reasons or about the outcome.	106	3.40%
Advise	Propose something or suggest doing something.	179	5.74%
Compromise	Reluctantly make a concession or acceptance out of necessity.	150	4.81%
Praise	Admire or speak highly of someone or something.	181	5.81%
Inform	Let someone know about something.	221	7.09%
Flaunt	Show off exaggeratedly in order to gain attention.	115	3.69%
Criticize	Censure or comment sharply on someone or something.	124	3.98%
Thank	Express gratitude for someone's help or kindness.	73	2.34%
Agree	Concur on something or with someone's viewpoint.	143	4.59%
Leave	Get away temporarily possibly to conclude the conversation.	109	3.50%
Query	Inquire others in order to find out something.	**311**	**9.97%**
Joke	Make exaggerated or humorous statements for entertainment.	165	5.29%
Overall	Sum of all intent labels.	**3118**	**100%**

Table 1: The quantity and proportion of each intent label with its brief description in dataset.

Dataset	Size	m	SA	IR
MDID(Kruk et al., 2019)	1299	t,v	✗	✓
MIntRec(Zhang et al., 2022)	2224	t,v,a	✗	✓
CH-SIMS(Yu et al., 2020)	2281	t,v,a	✓	✗
CSMAS(Ge et al., 2022)	1564	t,i	✓	✗
MSAIRS(ours)	3118	t,i	✓	✓

Table 2: Comparison of several multimodal datasets. *m* represents modalities. *t,v,a,i* represent text, video, audio, and image, respectively. *SA* and *IR* stand for sentiment analysis and intent recognition.

Modality	Sentiment	Quantity	Proportion
Context	Positive	728	23.35%
	Negative	921	29.54%
	Neutral	1469	47.11%
Sticker	Positive	1115	35.76%
	Negative	1166	37.40%
	Neutral	837	26.84%
Multimodal	Positive	1083	34.74%
	Negative	1358	43.55%
	Neutral	677	21.71%

Table 3: Statistics on the quantity and proportion of three sentiment labels in dataset.

and sticker. For pieces different from the manual annotation, we organize all annotators to discuss again and manually modify the labels. We find that the sentiment labels are almost entirely correct (96.8%), but a small part of intent labels (about 10%) are difficult for everyone to agree with GPT4, so we discard them. Finally after GPT4 review, we obtain 3118 pieces of data, the sentiment and intent labels of which are very convincing.

3.4 Data statistics

MSAIRS comprises 3.1k instances of data containing both context and stickers, along with 2.1k distinct stickers due to the pairing of the same sticker with different contexts. We divide the training set and the test set in a 9:1 ratio. In Table 2, we list the comparison between MSAIRS and several mainstream multimodal sentiment or intent datasets. As can be seen, while MSAIRS is larger in scale, it is the first dataset to include both sentiment analysis and intent recognition tasks. The quantity and proportion of multimodal intent labels is shown in Table 1, which closely aligns with the actual proportions found on social media. Table 3 presents the distribution of sentiment labels. We find that there are many samples where the sentiment labels obtained from different modalities are not consistent, which is demonstrated specifically in Figure 3. This indicates that relying solely on the context or sticker may not accurately determine the overall sentiment, emphasizing the importance of multimodal holistic analysis.

Our dataset also includes 70 instances where the same context paired with different stickers results in varying sentiment and intent labels, which is described in detail in Section 3.5. This illustrates the indispensable role of stickers in sentiment analysis and intent recognition on social media, underscoring the necessity of our research.

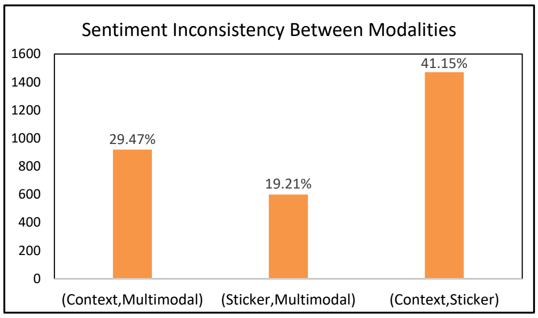

Figure 3: Inconsistency in sentiment labels across different modalities.

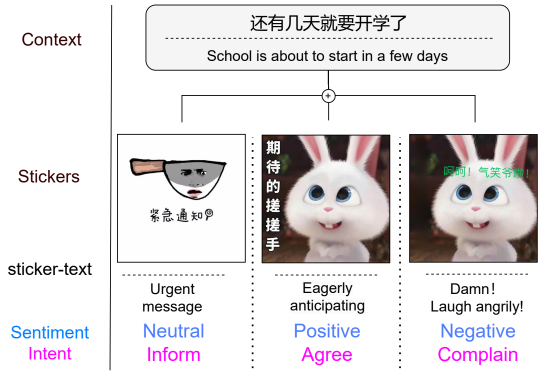

Figure 4: Examples from MSAIRS. It illustrates how the same context, accompanied by different stickers or sticker-texts, can convey entirely distinct sentiment and intent.

3.5 Illustrative example

Including the study of stickers can significantly contribute to the research on chat sentiment and intent. Figure 4 shows a set of examples extracted from our dataset. The context "School is about to start in a few days" conveys a neutral sentiment with several possible intent labels such as inform and complain, making it difficult to judge. The first sticker which expresses a neutral sentiment aims to inform others urgently. The rabbit in the second sticker depicts a joyful expression, signifying a positive sentiment and approval of the start of the school. Despite the visual similarity between the second and third sticker, the sticker-text "Damn! Laugh angrily!" reveals strong resistance and complaint about the start of school, forcing a smile negatively. From this, it can be seen that different stickers and sticker-texts can lead to completely different sentiments and intents.

Through specific examples, it can be observed that in order to recognize sentiment and intent in social media conversations, context, stickers, and sticker-text must be considered holistically. This further demonstrates the challenge and necessity of our task and dataset.

4 Model

4.1 Task Description

The objective of our baseline model MMSAIR is to predict both the multimodal sentiment label y_s and the multimodal intent label y_i with given context $X_t = (x_1, x_2, \ldots, x_m)$, sticker I_{img} and sticker-text $S_t = (s_1, s_2, \ldots, s_n)$. Here, x_i represents the i-th word in the context and s_i represents the i-th word in the sticker-text. m and n refer to the length of context and sticker-text respectively.

4.2 Encoding Layer

Text Encoder. BERT (Devlin et al., 2018) is a pre-trained natural language processing model that employs a masked language modeling training approach. In our study, we utilize two distinct, non-shared parameter BERT models to independently encode context and sticker-text. For classification tasks, we extract the hidden layer information of the first special token [CLS] in each sequence to serve as the representation for the entire sequence:

$$E_X = BERT(X_t),\ E_S = BERT(S_t). \quad (1)$$

Image Encoder. CLIP (Radford et al., 2021)is a multimodal pretraining model that establishes robust connections between images and text. We use CLIP as a image encoder to capture intrinsic image representations and reduce the disparity between text and images. To facilitate subsequent multimodal fusion, we further utilize Convolutional Neural Networks (CNN) for dimensionality reduction, ultimately obtaining the image representation E_I:

$$E_I = Conv1d(CLIP(I_{img})) \quad (2)$$

4.3 Multimodal Fusion Layer

First, in order to have a more comprehensive understanding of stickers, sticker and sticker-text should be input into the model as a whole. We initially concatenate the embeddings E_I extracted from the sticker images and the embeddings E_S from the sticker-texts, resulting in the composite embedding $E_{I,S}$:

$$E_{I,S} = Concat(E_I, E_S) \quad (3)$$

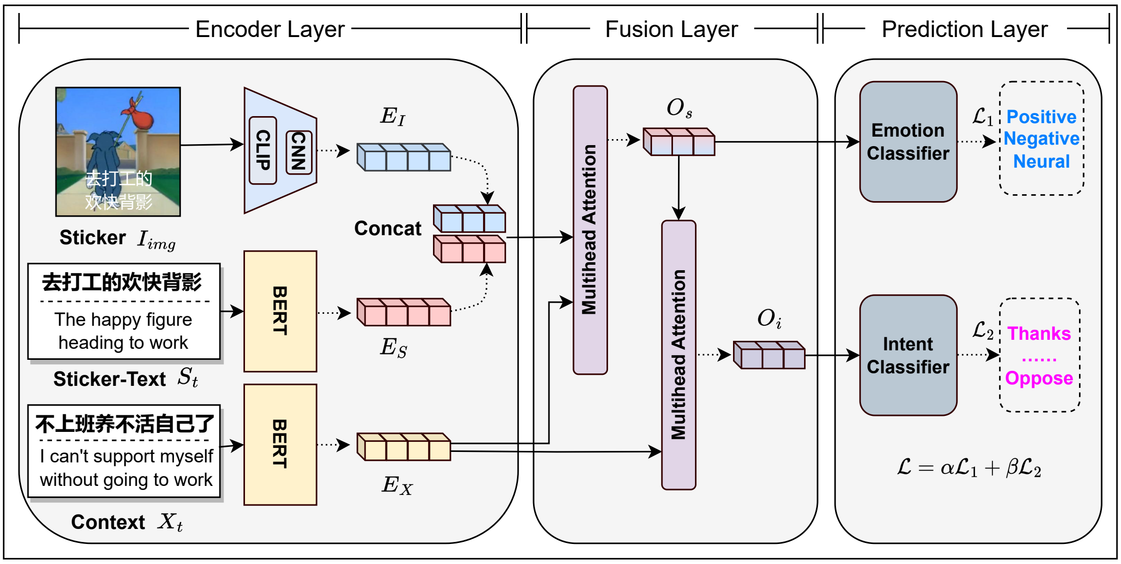

Figure 5: Abstract view of the multimodal baseline model MMSAIR.

Subsequently, to analyze the combined sentiment of the context and sticker, we input the composite embedding $E_{I,S}$ along with the context representations E_X into a multi-head attention (Vaswani et al., 2023) module for fusion, yielding the output O_s which will serve as a sentiment representation vector for sentiment classification in prediction layer:

$$O_s = MultiHead(E_X, E_{I,S}, E_{I,S}) \quad (4)$$

For intent labels, we find that intent differs from sentiments, as intent involves deeper semantic representations and certain intents exhibit distinct sentimental aspects. Therefore, in the second step, we input the shallow semantic embeddings O_s along with the text embeddings E_X into a multi-head attention module, resulting in the output O_i which will serve as an intent representation vector for intent classification in prediction layer:

$$O_i = MultiHead(O_s, E_X, E_X) \quad (5)$$

4.4 Prediction Layer

We predict the sentiment label using the output O_s obtained from the representation fusion layer. Initially, O_s is fed into a sentiment classifier composed of linear layers, followed by a softmax function for probability prediction. FFN stands for feed-forward neural network:

$$P(y_s|I_{img}, X_t, S_t) = softmax(FFN(O_s)) \quad (6)$$

Likewise, we pass O_i through an intent classifier and employ the softmax function for probability prediction:

$$P(y_i|I_{img}, X_t, S_t) = softmax(FFN(O_i)) \quad (7)$$

Our loss function consists of two components. $\mathcal{L}_1$ corresponds to the loss function for multimodal sentiment classification, while $\mathcal{L}_2$ represents the loss function for multimodal intent classification:

$$\mathcal{L}_1 = -\frac{1}{|D|}\sum_{|D|} logP(y_s|I_{img}, X_t, S_t) \quad (8)$$

$$\mathcal{L}_2 = -\frac{1}{|D|}\sum_{|D|} logP(y_i|I_{img}, X_t, S_t) \quad (9)$$

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 \quad (10)$$

where $|D|$ denotes all samples in the dataset. α and β are hyperparameters controlling the task weights. In our experiment both are set to 1, indicating that sentiment analysis and intent recognition are equally important.

5 Experiment

5.1 Experimental Setups

We compare our model with several popular unimodal and multimodal models. We use **BERT** (Devlin et al., 2018), **ALBERT** (Lan et al., 2019), and **RoBERTa** (Liu et al., 2019) as context-only baselines, using the context as input. For the image-only models, we utilize **ViT** (Dosovitskiy et al., 2020), **CLIP** (Radford et al., 2021), and **ResNet50** (He et al., 2016) as baselines, taking stickers as input. We set the size of ViT to 768 and use the ViT-based

	Models	Sentiment-Acc.	Sentiment-F1	Intent-Acc.	Intent-F1
Context-only	BERT	65.05	64.48	61.94	61.84
	ALBERT	60.55	60.00	44.98	44.48
	RoBERTa	66.78	66.55	65.74	65.40
Image-only	ViT	51.21	51.15	15.57	16.09
	Clip	58.13	57.59	21.11	20.11
	ResNet50	42.56	38.97	12.80	11.03
Multimodal	mBERT	68.86	**71.89**	65.05	58.40
	EF-CAPTrBERT	62.28	56.33	49.48	45.09
	PMF	65.95	62.27	52.60	47.01
	MMSAIR(ours)	**69.90**	69.43	69.82	**69.82**
Ablation Study	w/o C_F	63.67	63.91	10.73	7.36
	w/o S_F	68.51	68.39	67.13	67.01
	w/o ST_F	68.86	68.51	**69.89**	69.77
	w/o S_F&ST_F (Context-only)	38.41	31.03	68.51	68.39
	w/o C_F&ST_F (Image-only)	59.17	58.16	9.69	6.67

Table 4: Overall experimental results comparison. *Acc.* represents accuracy. *F1* represents the weighted F1 score. The *w/o* represents without. *C_F* represents context features. *S_F* represents sticker image features, and *ST_F* represents sticker-text features.

Task	Sentiment		Intent	
	Acc.	F1	Acc.	F1
SA	67.47	66.67	-	-
IR	-	-	68.17	67.93
MSAIRS	**69.90**	**69.43**	**69.82**	**69.82**

Table 5: Comparison of experimental results for Individual Sentiment Analysis, Intent Recognition Tasks, and the Joint Task of Both.

image encoder from CLIP. The same linear layer and classifier are added to unimodal models to obtain classification results. We choose **mBERT** (Yu and Jiang, 2019), **EF-CAPTrBERT** (Khan and Fu, 2021), and **PMF** (Li et al., 2023) as multimodal comparison models.

All models are trained on a NVIDIA GTX3090Ti for 200 epochs with the same parameters. We use the Adam (Kingma and Ba, 2014) optimizer, with a learning rate of 1e-5, weight decay of 1e-5, a training set batch size of 16, and a validation set batch size of 2.

5.2 Overall Results

Table 4 shows the experimental results of sentiment analysis and intent recognition on our MSAIRS dataset.

For context-only models, since they only consider the context, they could lead to one-sided results and insufficient performance. However, text information is relatively direct and easy to understand, and in many cases it can approach the final result, so the performance is not poor. For image-only models, not considering the sentimental and intentional tendencies of the text and only processing images which are abstract especially stickers, has led to poor performance, far lower than both multimodal and context-only models. Results of unimodal models show that the context contains richer sentiment and intent information compared to images. Especially for intent recognition tasks, it primarily relies on the context; solely relying on images can hardly make predictions. The importance of textual information further underscores the necessity to extract and process the sticker-text.

The performance of multimodal models is not bad. Among them, MMSAIR performs best in two joint tasks, which benefits from its proper handling of text and stickers and appropriate fusion. In the task of sentiment analysis, mBERT's performance is on par with MMSAIR, but its performance in intent recognition is still much lower compared to MMSAIR. This also shows that the MMSAIR model is simple but very effective, making it a strong multimodal baseline for MSAIRS dataset.

5.3 Multimodal Impact

As shown in Table 4, without the context, our model's performance in intent recognition decreases significantly. This indicates that context contains richer information about the intent, and considering only stickers is not sufficient for effectively recognizing the intent. When the image features or sticker-texts are removed, the result slightly decreases. It can be seen that both the images and sticker-texts can enhance predictive performance, which are indispensable for achieving the best results. Last two rows show the results of our multimodal model processing unimodal information when one modality is empty. In the absence of

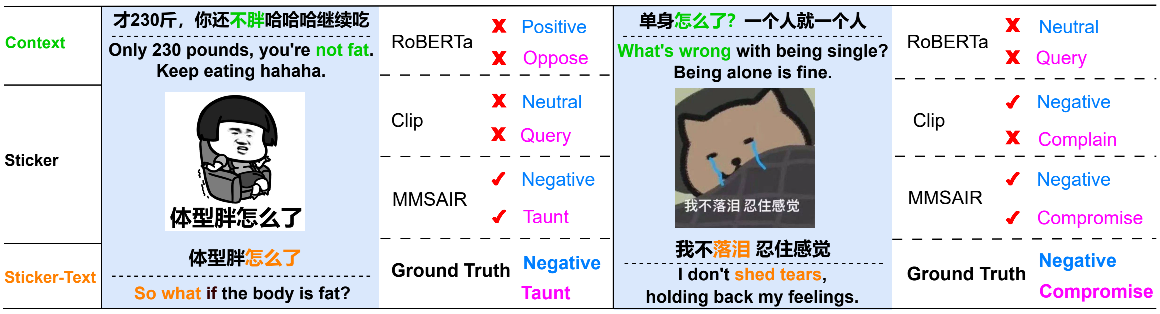

Figure 6: The results obtained by running typical examples in our dataset using different models.

S_F and ST_F, while our performance in intent recognition is good, the performance of sentiment analysis is quite poor, indicating that stickers play a significant role in expressing sentiments, sometimes even overshadowing the text. When C_F and ST_F are empty, corresponding to the second part of image-only models, our model's performance in sentiment analysis has improved a lot. Similar to other image-only models, the results in intent recognition is terrible, showing that images alone don't support intent recognition.

The results indicate that for the MSAIRS task, the three types of features, i.e., context, sticker, and sticker-text, are all necessary. Despite the challenging task, our model performs exceptionally well in handling multimodal information while also capable of handling unimodal data, demonstrating strong robustness.

5.4 Subtask Influence

Sentiment analysis and intent recognition tasks also influence each other. As shown in Table 5, the experimental results of each task alone on our dataset are poorer than those of the joint task proposed by MSAIRS. This indicates that the determination of sentiment has an important impact on recognizing specific intents and intents can also reflect sentiments, and it also demonstrates the effectiveness of performing the two tasks jointly, suggesting that the MSAIRS task is valuable.

6 Case Study

In Figure 6, we show two typical examples from our dataset and their experimental results on the best context-only RoBERTa model, the best image-only Clip model, and our MMSAIR model.

In the first example, the context-only model focuses on the text of “you're not fat" and “hahaha", providing an opposite intent and positive sentiment. The image-only model only processes the sticker, where the character appears indifferent, and there is sticker-text “so what?", so it concluded with neutral and querying. MMSAIR combines the context and sticker, seeing that although the text says “not fat", in reality, the “hahaha" is actually making fun of the other person. The sticker also conveys the irony, so it should be negative and taunting. In the second example, due to the context being a general question, the context-only model judges this as a neutral query. The cartoon character in the sticker is crying and has the text “shed tears", so the image-only model believes this is a negative complaint. However, by combining the context and sticker, it is found that the speaker actually expresses helplessness about being single in the context, using the sticker's crying to express negative sentiment. Therefore, the conclusion of MMSAIR is negative and compromising, same as the Ground Truth. Even the best unimodal models find it difficult to correctly predict the sentiment and intent of chat records with sticker on social media. However, MMSAIR is able to judge the overall sentiment and intent even for more difficult examples in the experiment, indicating its effectiveness.

7 Conclusion

We study the impact of stickers in social media on sentiment analysis and intent recognition. We propose MSAIRS task, a manually annotated dataset along with thorough review, and a multimodal baseline MMSAIR. The low performance of other models further proves the challenge of our task and dataset, demonstrating the necessity to integrate visual information from stickers. In the future, it is worth exploring how to better fusion these multimodal information to better solve this task or add new useful modalities.

Limitations

The chat records and stickers used in this paper are all sourced from Chinese social media platforms. A small portion of the data may exhibit differences in understanding and expression due to disparities in Chinese and Western cultures.

References

Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Jean H French. 2017. Image-based memes as sentiment predictors. In *2017 International Conference on Information Society (i-Society)*, pages 80–85. IEEE.

Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2019. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

Feng Ge, Weizhao Li, Haopeng Ren, and Yi Cai. 2022. Towards exploiting sticker for multimodal sentiment analysis in social media: A new dataset and baseline. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6795–6804, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xuejian Huang, Tinghuai Ma, Li Jia, Yuanjian Zhang, Huan Rong, and Najla Alnabhan. 2023. An effective multimodal representation and fusion method for multimodal intent recognition. *Neurocomputing*, page 126373.

Constance Iloh. 2021. Do it for the culture: The case for memes in qualitative research:. *International Journal of Qualitative Methods*, 20(2):153–163.

Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Widia Lestari. 2019. Irony analysis of memes on instagram social media. *PIONEER: Journal of Language and Literature*, 10(2):114–123.

Marc D Lewis et al. 2005. Getting emotional: A neural perspective on emotion, intention, and consciousness. *Journal of Consciousness Studies*, 12(8-9):210–235.

Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. 2023. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613.

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

T Nikil Prakash and A Aloysius. 2021. Hybrid approaches based emotion detection in memes sentiment analysis. *International Journal of Engineering Research and Technology*, 14(2):151–155.

Raj Ratn Pranesh and Ambesh Shekhar. 2020. Memesem: a multi-modal framework for sentimental analysis of meme via transfer learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*.

Hemant Purohit, Guozhu Dong, Valerie Shalin, Krishnaprasad Thirunarayan, and Amit Sheth. 2015. Intent classification of short-text on social media. In *2015 ieee international conference on smart city/socialcom/sustaincom (smartcity)*, pages 222–228. IEEE.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the evolution of (hateful) memes by means of multimodal contrastive learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 293–310. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. Learning transferable visual models from natural language supervision.

Shiyue Rong, Weisheng Wang, Umme Ayda Mannan, Eduardo Santana De Almeida, Shurui Zhou, and Iftekhar Ahmed. 2022. An empirical study of emoji use in software development communication. *Information and software technology*.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13:277–289.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93.

Ying Tang and Khe Foon Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International Journal of Communication*, 13:27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2887–2899, New York, NY, USA. Association for Computing Machinery.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

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.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697.

Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. 2021. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042.