I Mean I Am a Mouse: Mmeets for Bilingual Multimodal Meme Sarcasm Classification from Large Language Models

yunzhe Liu* 2101010410@stu.wit.edu.cn
Wuhan Institute of Technology
xinyi Xu* xuxinyii0307@163.com
Chengdu University of Technology

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Multimodal image-text memes are widely used on social networks and present significant challenges for high-precision sentiment analysis, social network analysis, and understanding diverse user communities, especially due to their deep cultural and regional influences. However, most existing studies on multimodal memes focus primarily on English-speaking communities and on preliminary tasks, such as harmful meme detection. In this paper, we focus on a more specific challenge: high-precision sarcasm classification in various contexts. We introduce a novel dataset for classifying sarcasm in multimodal memes, covering both Chinese and English languages. This dataset serves as a critical resource for developing and evaluating models that detect sarcasm across different cultural contexts. Furthermore, we propose a framework named Mmeets, which leverages Large Language Models (LLMs) and abductive reasoning to interpret the relationships between images and text, enhancing text understanding. Mmeets employs a pre-trained AltCLIP vision-language model alongside a cross-attention mechanism to effectively fuse image and text data, capturing subtle semantic connections. Our experimental results show that the Mmeets method outperforms state-of-the-art techniques in sarcasm classification tasks.

Keywords: Multimodal learning ;Large language model ; Sentiment analysis.

1. Introduction

Multimodal image-text combination memes have become ubiquitous on social networks and have become a key way for users to express their emotions and opinions. These memes not only reflect deep cultural connotations and geographical characteristics, but also provide an important basis for understanding users' emotions and intentions. From enhancing user interaction to monitoring the health of society, these memes play an important role in several application domains. By analyzing the image and text information in emojis, we can better capture the emotional changes of users and the dynamics of social opinions, providing effective support for personalized recommendations, brand marketing, and public opinion monitoring.

Although extensive memes sentiment analysis has been conducted in the English-speaking community, there is still a significant research gap in terms of comprehensive analysis across different languages and cultures. Existing research has mainly focused on sentiment tasks

* Both authors contributed equally to this research.

such as detecting harmful memes and recognizing emotions, often neglecting the subtle interactions between text and images that convey complex emotions such as sarcasm. Sarcastic memes express complex emotions and critical perspectives by humorously reinterpreting elements of popular culture, social phenomena, or everyday life. Therefore, more comprehensive sentiment analysis studies in specific cultural contexts are needed to capture the deeper emotions and intentions embedded in these emojis in order to enhance the accuracy and effectiveness of user interaction experience and social health monitoring.

Compared to the detection of mockery in multimodal memes, classifying sarcasm within them is more challenging. Self-mockery is a form of mockery in which a person jokes about his or her own behavior, traits or situation through humor or sarcasm this form of mockery is a self-reflective expression usually aimed at demonstrating self-criticism or relieving embarrassment and stress. In contrast, mocking others uses sarcasm to criticize other people, groups or situations. Self-mockery and ridiculing others often involve complex semantic interpretations and deep cultural contexts, which makes it difficult for models to understand and decode these subtle linguistic differences and cultural connotations. The expression of sarcasm not only requires the model to have advanced semantic comprehension, but also requires it to be culturally sensitive in order to accurately capture and differentiate these complex emotional expressions. In addition, the nuances of self-mocking and mocking others may be interpreted differently in different cultures and contexts, which further increases the difficulty of the analysis. Therefore, when performing sentiment analysis in multilingual and multicultural contexts, the model must be able to handle and understand more complex semantic and cultural contexts to ensure accurate recognition of self-moking and mocking others in multimodal memes.

In order to address the challenge of high-precision sentiment classification when conveying complex emotions such as through textual and image multimodal memes in cross-linguistic and cultural contexts, this paper introduces a bilingual multimodal meme sarcasm classification dataset and proposes an approach called Mmeets. This approach is inspired by the success of Large Language Models (LLMs) in using background knowledge for reasoning at the cognitive level. Firstly, we utilise LLMs for causal reasoning in order to enhance the multimodal capabilities of the models, enabling them to create multimodal representations within the text embedding space for the purpose of complementing textual contextual information and enhancing the text. Second, we separate image and text features and adapt the two embedding spaces to specific downstream tasks. We use the pre-trained AltCLIP visual-linguistic model to achieve effective multimodal information extraction. Finally, we achieve multimodal fusion through a cross-attention mechanism and classify multimodal emoticons as mocking others or self-deprecating. This approach enables better understanding and parsing of complex emotional expressions in emoticons, and improves the accuracy and reliability of emotion classification. Our contributions are as follows:

1. We introduce a bilingual (Chinese and English) multimodal meme sarcasm classification dataset, which has been annotated and verified by experts.

2. A novel approach called Mmeets for multimodal meme sarcasm classification is proposed,which leverages aductive reasoning with LLMs with a frozen pre-trained AltCLIP vision-language model.

3. To our best knowledge, we are the first to alleviate the issue of superficial understanding for sarcasm meme classification by explicitly utilizing commonsense knowledge, from a fresh perspective on harnessing advanced LLMs.

2. Related Works

2.1. Meme Sentiment Analysis

Sentiment analysis is a natural language processing technique designed to identify and extract subjective information from text, thereby determining its sentiment tendency. In early research, the focus was primarily on the sentiment analysis of text-based social interactions, such as social media posts and e-commerce platform product reviews. This was used to uncover the deeper meanings expressed by individuals in these conversations. Examples include ternary sentiment analysis of multidomain comments Xu et al. (2020)and analyzing the emotions conveyed by YouTube videos through user comments Asghar et al. (2015).With the development of deep learning techniques there has also been a gradual emergence of sentiment analysis of images You et al. (2015); Yuan et al. (2013) and speech Lakomkin et al. (2019); S and M (2018).

As social media continues to evolve, contemporary social dialogues increasingly encompass multimodal information, such as the combination of text and images, rather than being limited to single-modal content. Meme sentiment classification exemplifies multimodal sentiment analysis, as memes comprise both textual and visual information. Investigating the interaction between these two modalities presents a significant challenge. Presently, most research on meme sentiment analysis concentrates on detecting harmful Burbi et al. (2023); Lin et al. (2023) and negative memes He et al. (2016b). These sentiment classifications typically have a strong degree of distinction. However, there is a lack of research on the detection of sentiments like sarcasm, which are characterized by polysemy, metaphor, and strong contextual dependence.

2.2. Sarcasm meme classification

Sarcasm memes frequently appear in social media posts, chat messages, and various other forms of social interaction. They often carry strong elements of aggression and insult, accompanied by cyberbullying and hate speech. Regular exposure to sarcastic memes can have negative effects on users' mental health, particularly for adolescents and other impressionable groups. Prolonged exposure to such content may lead to psychological trauma.In addition, sarcasm memes may contain hate speech such as sexism, racial discrimination, religious discrimination, etc., which may intensify social conflicts and clashes and undermine social harmony and stability. Therefore, it is particularly important to classify and detect sarcastic memes.

In recent years, driven by advancements in deep learning, many studies have devoted to multimodal sarcasm meme detection, some approaches pursue the aligning of visual and textual representation vectors Qin et al. (2023), while others consider how to better capture the emotions conveyed by images Hee et al. (2024). However, most of the current studies have been devoted to the detection of sarcastic meme in a single language and have only classified sarcastic and non-sarcastic messages. In social media, self-mocking

memes also appear frequently. These memes express humor and self-mocking to create a light-hearted atmosphere and generally do not have negative impacts on others. However, because the emotional expressions of self-mocking memes and mocking others memes are not significantly different , many sarcasm meme detection methods are likely to classify self-mocking as a type of mocking others as well. Due to the nuances between self-mocking and mocking others memes, we need a deeper understanding of meanings interwoven behind visual and textual information. This is a significant challenge, and to the best of our knowledge, there is currently no research addressing this issue.

The methods closest to our work is ISSUES Burbi et al. (2023), which investigate the fusion of visual and textual information to understand metaphorical information. On this basis, we further investigate the subtle difference in meaning between self-mocking and mocking others, and extend monolingual sarcasm meme classification to bilingual sarcasm meme classfication in English and Chinese.

2.3. Large Language Models

With the development of Large Language models, more and more people use Large Language models for multimodal sarcastic classification, such as multimodal sarcasm detection using Large Language models combined with visual guidance and exemplar retrieval Tang et al. (2024) and multimodal inference using Large Language Models Yin et al. (2023). One of the similar to our work is abductive inference through Large Language Models Lin et al. (2023), where abductive reasoning means that the Large Language Models are able to perform deep inference on meme texts to identify the implicit meaning and underlying sentiment in them.

LLMs have been pivotal in advancing how machines comprehend complex human emotions and subtleties embedded in diverse formats of data, including text and images.Recent advancements have extended the utility of LLMs beyond mere text analysis, integrating them into multimodal contexts where they analyze and interpret the interplay between textual and visual data. This integration is particularly crucial in the analysis of multimodal memes, where the meaning often emerges from the nuanced combination of text and imagery. Models such as OpenAI's GPT series and Google's BERT have been adapted to better handle such data through techniques that combine natural language processing with image recognition capabilities.

Recently, LLMs have demonstrated remarkable capability in complex reasoning such as generating intermediate inference procedures before the final output. Traditional models have often struggled with the implicit meanings and cultural nuances that sarcasm entails, especially when it crosses linguistic and cultural boundaries. The employment of LLMs facilitates a deeper understanding of these complexities by leveraging vast amounts of data and sophisticated inferencing capabilities.

In this work, we conduct abductive reasoning with LLMs, which further advocates a multimodal reasoning paradigm to enhance with strategetic text prompting for sarcasm meme classification.

3. Dataset

3.1. Overview

Currently, known sarcasm meme detection datasets mainly target the classification of sarcasm and non- sarcasm, meme, and most of the relevant datasets are limited to the English sarcasm meme classification dataset, and there does not exist a Chinese-English dataset with a more nuanced distinction between self-mocking and mocking others. In order to better explore the difference between self-mocking and mocking others, we created the first bilingual self-mocking and mocking others dataset in English and Chinese: BSMM. We collected more than 5000 relevant data from various social media, which contains 2500 datasets each in English and Chinese, and its specific distribution is shown in Table 1. Figure 1 shows some examples from our dataset.They contain a meme image, a piece of text obtained through ocr recognition and labels that have been labelled and corrected by scholars and experts.

BSMM datasets	Chinese	English	Total
mock others	1299	1328	2627
Deprecating	1278	1378	2656
total	2577	2706	5283

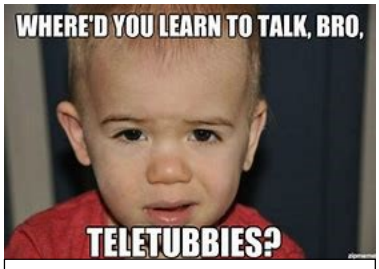
Table 1: Class distribution of meme dataset

3.2. Data Collection

In order to collect the meme of self-mocking and mocking others, we crawl the mainstream social media.For the English data, we mainly searched for keywords such as mockong, self-mocking, and meme, and crawled more than 3000 meme images from well-known social platforms such as Facebook, X, Reddit, etc. For the Chinese dataset, we mainly crawled Baidu Post Bar, Zhihu, Weibo, etc, and crawled more than 3000 meme images from these platforms. In the end, we crawled more than 6000 relevant meme images. but we found that more than 90% of the memes did not have labels, and even if labels existed for the memes, the labels might not be classified correctly due to a variety of reasons. The accuracy of the meme labels was crucial to our subsequent study, so we asked 10 linguistic and 10 psychological scholars to label all the data.Since self-mocking and mocking others are expressed in a similar way, it is difficult to accurately capture the subtle differences, so in order to ensure the correctness of the labelling, we gave the same meme to three different scholars for labelling, if the three labelled the same result we considered the labelling as accuracy,If they are not same we submit them to linguistic and psychological experts for further validation and re-labelling. After labeling and correcting by scholars and experts, we obtained nearly 4000 sheets of high quality meme data. And the text in it was extracted.

Meme text:我钱呢。。鼠鼠啊‥我钱呢
Label:self-mocking

Meme text:求你了给你脑子排排水吧
Label:mocking



Memetext: WHERE'D YOU LEARN TO TALK,BRO,TELETUBBIES?
Label: moking

Meme text: Me Sad Music My problems
Label:self-moking

Figure 1: Presentation of some Chinese and English meme data

4. Our Approach

4.1. Definition of the problem

We define our datasets in two parts, Chinese and English datasets, and define the Chinese meme dataset as a set of memes, each of which is defined as a ternary such as $M_{ch} = \{I, T, y\}$.Similarly the English meme dataset is defined as $M_{en} = \{I, T, y\}$ where I is the meme image, T is the meme text, and y is the meme's label $y \in$ {self-mocking, mocking others}. We take image I, text T as input and label y as output. It is used to complete the mocking others meme classification work.

4.2. Abductive Reasoning with LLMs

Large Language Models are rich in knowledge and thinking ability, and they can effectively integrate multimodal information to achieve comprehensive understanding and classification of complex content. So more and more people are using Large Language Models for multimodal sarcasm classification.Through the powerful reasoning and ability to capture details in Large Language Models, they often achieve excellent results.

Abductive reasoning Lin et al. (2023) is: a method of logical reasoning aimed at deducing possible causes from phenomena. Compared with deductive and inductive reasoning, abductive reasoning focuses more on explanation and hypothesis formation, and is particularly suitable for reasoning tasks with incomplete information. Large Language Models

can play an important role in abductive reasoning through their powerful natural language processing capabilities and learning from large-scale data.It can make full use of the known information to deduce more hidden details. With Laege Language Model, it is possible to fully understand image and textual information. Compared to other methods, it is effective to understand the information generated by multiple modal interactions, capture subtle differences between different data, and thus infer the reasons for the results.For multimodal data such as images and text with few textual information, the abductive reasoning can capture the correlation information between images and text to supplement the context of the text, so as to achieve the effect of competing for strong text.

4.3. ALTCLIP

ALTCLIP Chen et al. (2023) is a powerful Chinese-English bilingual pretraining model improved based on CLIP Radford et al. (2021). The traditional CLIP is a model developed by OpenAI that understands the relationship between images and text, and achieves image and text alignment by training on large-scale image pair data. it is a contrastive learning model, which aligns image information and text information through the method of contrastive learning, and is able to find out the correlation between text and images well. It has achieved excellent performance in the downstream tasks. Bachard and Maugey (2024)However, the mainstream exploration of CLIP focuses on English datasets, and there is a lack of Chinese datasets, and the effect of CLIP is not ideal for the downstream task of Chinese data, based on this reason there is the ATCLIP model.Designed for bilingual environments, ALTCLIP is able to effectively process and understand both English and Chinese text and corresponding images. This makes ALTCLIP more widely used in multilingual environments, especially for cross-cultural data analysis and processing involving both Chinese and English languages. ALTCLIP utilizes a large number of Chinese and English images to pre-train the data. These data not only contain images from different domains and scenarios, but also cover a rich variety of linguistic expressions, thus enhancing the performance of the model in different application scenarios. The ALTCLIP model is divided into two phases. The first phase use the freezing of the CLIP text encoder and the teacher model. we Fine-tune the XLMR model on Chinese and English text datasets through teacher modeling to enable XLMR to learn the capabilities of the CLIP text encoder. In the second stage, the text encoder in the CLIP model is replaced by XLMR, and the CLIP imgae encoder is frozen, and the XLMR model is further trained with the Chinese and English text datas and the image datas, so that the text representational features can be better aligned with the visual representational features.

4.4. Mmeets

Figure 2 shows an overview of our approach.Mmeets mainly includes three areas:(1)Enhancing Textual Representation;(2)Text-Image Information Alignment;(3)Sarcastic Meme Inference.

4.4.1. Enhancing Textual Representation

In this paper, we first extract the text caption T of the image I by off-the-shelf captioning models Mokady et al. (2021) , and then write a text prompt template S that is handed over to the LLMs for processing.After after handing over the ternary $\{I, S, y\}$ to the LLMs as

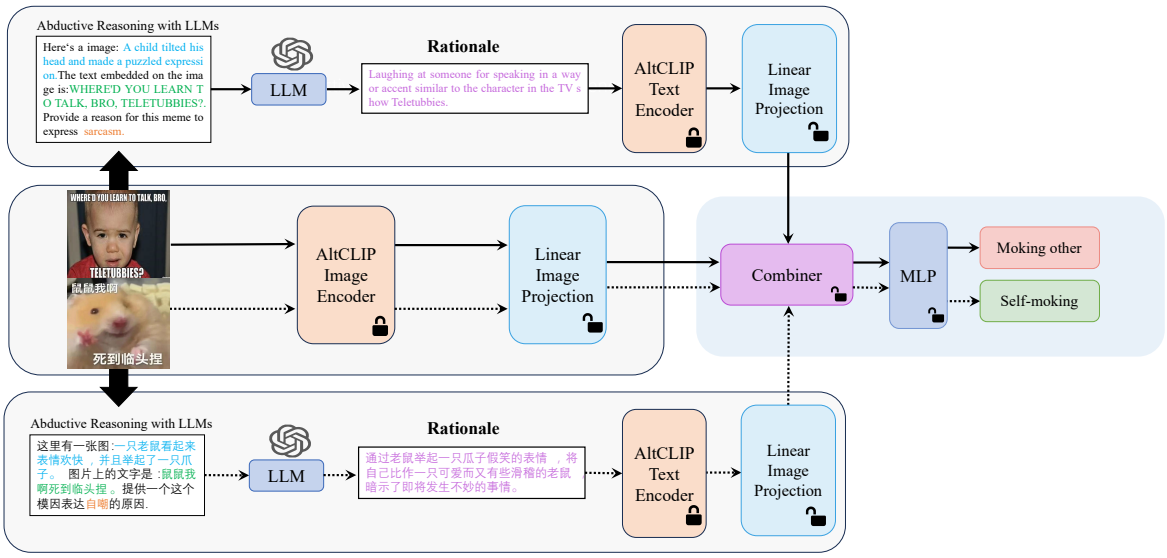


Figure 2: Overview of the proposed approach. We do abductive inference on the Chinese and English datasets separately, then use the ALTCLIP model with frozen parameters to extract the information of images and texts, and finally use the cross-attention mechanism for modal combination.

the basic element to make it carry out abductive reasoning to get the principle R of meme labelled as the self-mocking or mocking others.The basic structure of the text template S is:

English:If you are given a image [I] with the text [T] on it, combined with the image and the text, please answer how does it express [y]? Reply to the message within 50 words and use English.

Chinese:给你一张图片[I]，其上面存在文字[T]，结合图片和文字，请回答其是如何表达[y]的呢？回复信息在50字之内,请用中文回复。

We give the prompt templates in English and Chinese respectively, but there are a few points worth noting here:(1)We give the result of the final meme classification directly in the prompt template, hoping that the LLMs can reason out the intermediate process, which is the key to the whole abductive inference, and its by this method that we hope to solve the illusion of the LLMs to a certain extent, so that the thinking process given by it is correct.(2) We have strictly controlled the number of words in the template to be 50, the reason for doing so is that this ensures that the LLMs will not generate a long, redundant explanation process, avoiding the subsequent huge computation, and at the same time not let the generated textual information be too short, resulting in the inability to adequately explain the cause of the classification of meme.

4.4.2. Text-Image Information alignment

Text and image information alignment is the key step to mutimodel sarcasm meme detection. we mainly use ALTCLIP text encoder and ALTCLIP Image encoder to accomplish information extractio and alignment.

Specifically,We freeze the visual encoder $V(x)$ and the text encoder $T(x)$ in ALTCLIP its former use VIT Dosovitskiy et al. (2021), and the latter use XML-R Conneau et al. (2020).For a meme sample M we first encoder image I and its corresponding text R through enhancing textual to gain their embedding vectors as follows:

$$E_V = V(I) \tag{1}$$

$$E_T = T(R) \tag{2}$$

where E_V is visual embedding vectors,E_T is text embedding vectors.Then the obtained image embedding vectors and text through enhancing textual embedding vector are projected to the same d-dimensional space by linear image projection and linear text projection respectively, which is convenient for the subsequent extraction of fusion information of image and text information, and its specific calculation process is as follows:

$$E_{vd} = W_V E_V + b_V \tag{3}$$

$$E_{td} = W_T E_T + b_T \tag{4}$$

where $E_{vd} \in \mathbb{R}^{m \times d}$ is visual embedding vectors through linear image projection,$E_{td} \in \mathbb{R}^{m \times d}$ is text embedding vectors through linear image projection.

4.4.3. Sarcasm meme Inference

After going through the first stage of text enhancement and the second stage of information alignment, we obtain text and image vectors. Next, to enhance the semantic alignment between text and image for improved contextual understanding, we utilize a cross-attention mechanism for the multimodal fusion of textual and visual information:

$$Q_V = W_Q E_{vd} + b_Q, \tag{5}$$

$$K_T = W_K E_{td} + b_K \tag{6}$$

$$V_T = W_V E_{td} + b_V \tag{7}$$

$$E_o = \text{softmax}\left(\frac{Q_V K_T^\top}{\sqrt{d_k}}\right) V_T, \tag{8}$$

$$y = \text{softmax}(W_o E_o + b), \tag{9}$$

where E_o is Multi-modal fusion output, W_o denotes the linear projection, b is the bias. Through the cross-attention mechanism, we effectively capture the hidden information of text and image fusion. We then complete the final sarcasm meme classification with a MLP. The loss function L for the training data D is defined as:

$$L = -\sum_{(x,y)\in D} \log p(y|x). \tag{10}$$

which calculates the cross-entropy loss between the classified labels and the true sarcasm label y. Through the sarcasm meme inference stage, we effectively capture the information of image-text crossover, and finally complete the evaluation of the model using the cross-entropy loss function.

5. Experiment

5.1. Comparison Models

We compare Mmeets to several state-of-the-art sarcastic meme detection models, including single-modal models :(1)Text BERT Devlin et al. (2019); (2)ResNet He et al. (2016a). multimodal model: (3)VisualBERT Li et al. (2019); (4)CLIP Radford et al. (2021); (5)BLIP Li et al. (2022); (6)ISSUES Burbi et al. (2023). We used accuracy and macro average F1 scores as evaluation metrics.

Dataset	English		Chinese	
Model	Accuracy	Macro-F1	Accuracy	Macro-F1
Text BERT	69.57	69.56	66.67	66.67
ResNet	76.92	76.60	72.00	73.44
VisualBERT	79.00	79.13	77.81	76.41
CLIP	82.81	82.81	73.75	76.25
BLIP	81.60	81.60	80.63	80.75
ISSUES	87.60	87.62	80.62	80.62
GPT-4o	73.22	75.23	76.18	79.53
Mmeets	**88.80**	**88.84**	**90.03**	**90.01**

Table 2: Performance comparison of different models on English and Chinese datasets.The accuracy and macro-averaged F1 score (%) are reported as the metrics. The best results are in bold.

Table 2 shows the performance of our proposed method Mmeets and other comparison methods on both Chinese and English sarcasm datasets. According to the observations: (1) The first set of baseline models used only unimodal models, which extracted features from only a single text or image. Since self-mocking, mocking others modalities are nuanced,metaphorical, and there are cases where the text and images have opposite meanings, it is difficult for a single modal message to capture its hidden meanings, so its performance is poor. (2) The baseline models of the second group are all multimodal models, and the performance of the multi-modal model is significantly better than that of the unimodal model. CLIP and BLIP models use the method of contrast learning to further capture the association information between images and text, and the effect is better than the traditional multi-modal model VisualBERT. (3) The ISSUES model uses text inversion technology to further strengthen the alignment of text features and image features, making it perform best in the second group. However, through observation, it can be found that the performance of the second group of models on Chinese datasets is unsatisfactory. This

is due to the fact that they are all focus on English datasets and do not well in Chinese datasets.

The accuracy of our proposed model based on the best baselien model is improved by 1.22% and 9.35% on the English and Chinese datasets, We observe that: (1) on the English dataset, we have improved compared to all other models, which indicates that the abductive inference method we adopt can capture the associated content and the associated mode between images and texts, and make up for the lack of context in the textual information, so that the textual information has been sufficiently enhanced. And the enhanced text information is fused with the image information through the cross-attention mechanism to capture the information that is not effectively captured in the abductive inference stage; (2) On the Chinese dataset, we have a huge improvement compared to all the other models, which shows that ALTCLIP text encoder effectively acquires the capability of CLIP and extends it to Chinese data processing.

5.2. Ablative Study

We examine ablative studies on several variants of Mmeets in detail: (1) w/o reasoning: just fine-tune the altclip model with self-mocking inference, and no LLMS to enhance the text data. (2) w/o Visual Features: the visual features of the model are discarded, and only the original text features in the meme are used. (3)w/o Text Features: Discard the textual features of the meme and only use the visual features in the meme.

Dataset	**English**		**Chinese**	
Model	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**
Mmeets	88.80	88.84	90.03	90.01
w/o Reasoning inference	84.40	84.65	82.81	83.48
w/o Visual Features	68.00	65.12	66.56	73.18
w/o Text Features	79.12	78.69	74.37	74.05

Table 3: Ablation studies on our proposed framework.

As shown in Table 3: Compared to the Mmeets ablative modeling metrics are down in all respects. Specifically, w/o reasoning inference, there is a significant decline in both accurate and Macro-F1, a lack of Abductive reasoning, an inability to capture the relationship between parts of the text and images, and an inability to contextualize the text. w/o Visual Features and w/o Text Features, there is a huge drop in accuracy and Macro-F1 metrics, which suggests that for a fine-grained classification task such as mocking others and self-mocking memes, relying on a single modality's information is finite and does not allow for a full understanding of the metaphors in the meme. In contrast to this, multimodal approach, both text and image information is obtained, and the expansion of the amount of information allows for a better classification of memes. Here's an interesting note: although both are unimodal, w/o Visual Features are far less effective than w/o Text Features. We speculate that this is greatly related to mocking others, and the self-mocking meme has the metaphor, because the amount of textual information in the meme is very small, and its context is missing, there may be a single textual message that expresses a mocking

other person's meaning, but combined with a image message that becomes expressive of self-mocking.

5.3. Visualisation of the text enhancement process

Notice that the abductive inference through the LLMs, although its intermediate inference ratio is not the final output of the mock modal classification, contextualises the text to achieve textual enhancement. And the performance is excellent. We will start with a small number of samples to understand the process and details of our proposed Mmeets model in performing retrocausal inference in a more transparent and intuitive way. As shown in Figure 3.

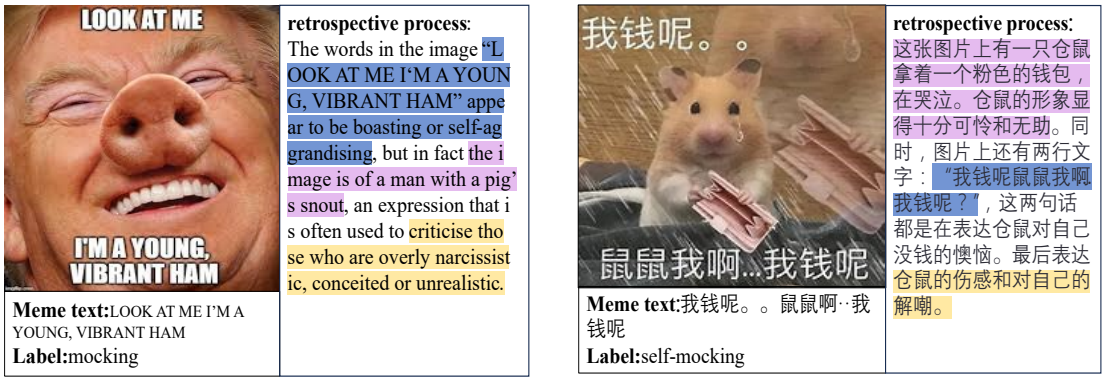


Figure 3: Examples of the process of abductive inference of the Chinese and English dataset meme

Visualising the intermediate process of abductive inference, we find that the LLMs inference as follows: (1) give the content represented by the text as the blue part of the figure, (2) give the content represented by the image as the pink part of the figure. (3) fuse the content represented by the text and the image for the final inference.

For the specific details of abductive inference, we explain through the example in Figure 3. for the English meme.It first analyses the meaning of the text in its meme: "LOOK AT ME I'M A YOUNG", the LLMs thinks that this sentence expresses the meaning of bragging, it is true that as over just look at this sentence alone may feel that it is someone who is narcissistic, and it does not express the meaning of mocking, but the LLMs then analyses the content of the image, which is a person but its has a pig nose. The content of the image is a person but it has a pig's nose. The LLMs combines the information from the text and the image and concludes that this is a jibe at people who are too narcissistic, conceited or unrealistic.For the Chinese meme, the LLMs first analyses its image: a hamster with tears in his eyes and an empty pink wallet in his hand, which expresses the hamster's pitifulness and helplessness, and then analyses the text: "Where's my money, rat me ah... Where's my money?", which expresses the hamster's chagrin at his lack of money, and finally combines the text and image information to conclude that this meme expresses the hamster's sadness and his self-mocking.

6. Conclusions

In this paper, we introduce BSMM, the bilingual (English and Chinese) dataset dedicated to mocking and self-mocking memes.Additionally,we propose Mmeets, a novel method for sarcasm meme classification, designed to effectively discern the nuanced meanings embedded within the interplay of images and their accompanying texts. Our approach begins with abductive reasoning utilizing large language models (LLMs) to enrich the textual context, thereby augmenting the expressive power of the text. We then employ the pre-trained ALTCLIP model, freezing its image and text encoders to extract respective content. Subsequently, we implement a cross-attention mechanism to merge the image and text information, enabling the model to capture the underlying subtleties of the memes. Comparative evaluations demonstrate that our model achieves state-of-the-art (SOTA) results on the BSMM dataset.

For future work, given the difficulty in assessing the quality of intermediate reasoning, where evaluation is inherently qualitative, we plan to conduct a systematic study on explainable harmful meme detection. This study will aim to establish explainability through evaluations involving human subjects.

References

Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, and Fazal Masood Kundi. Sentiment analysis on youtube: A brief survey. *CoRR*, abs/1511.09142, 2015. URL http://arxiv.org/abs/1511.09142.

Tom Bachard and Thomas Maugey. Can image compression rely on clip? *IEEE Access*, 12: 78922–78938, 2024. doi: 10.1109/ACCESS.2024.3408651. URL https://doi.org/10.1109/ACCESS.2024.3408651.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Mapping memes to words for multimodal hateful meme classification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 2824–2828. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00303. URL https://doi.org/10.1109/ICCVW60793.2023.00303.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in CLIP for extended language capabilities. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8666–8682. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.552. URL https://doi.org/10.18653/v1/2023.findings-acl.552.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10,*

2020, pages 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.747. URL https://doi.org/10.18653/v1/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016a. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Saike He, Xiaolong Zheng, Jiaojiao Wang, Zhijun Chang, Yin Luo, and Daniel Zeng. Meme extraction and tracing in crisis events. In *IEEE Conference on Intelligence and Security Informatics, ISI 2016, Tucson, AZ, USA, September 28-30, 2016*, pages 61–66. IEEE, 2016b. doi: 10.1109/ISI.2016.7745444. URL https://doi.org/10.1109/ISI.2016.7745444.

Ming Shan Hee, Rui Cao, Tanmoy Chakraborty, and Roy Ka-Wei Lee. Understanding (dark) humour with internet meme analysis. In Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, editors, *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 1276–1279. ACM, 2024. doi: 10.1145/3589335.3641249. URL https://doi.org/10.1145/3589335.3641249.

Egor Lakomkin, Mohammad-Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. Incorporating end-to-end speech recognition models for sentiment analysis. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 7976–7982. IEEE, 2019. doi: 10.1109/ICRA.2019.8794468. URL https://doi.org/10.1109/ICRA.2019.8794468.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning*

Research, pages 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. URL http://arxiv.org/abs/1908.03557.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9114–9128. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.611. URL https://doi.org/10.18653/v1/2023.findings-emnlp.611.

Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021. URL https://arxiv.org/abs/2111.09734.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. MMSD2.0: towards a reliable multi-modal sarcasm detection system. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10834–10845. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.689. URL https://doi.org/10.18653/v1/2023.findings-acl.689.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Maghilnan S and Rajesh Kumar M. Sentiment analysis on speaker specific speech data. *CoRR*, abs/1802.06209, 2018. URL http://arxiv.org/abs/1802.06209.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.97.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. Position-aware tagging for aspect sentiment triplet extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2020, Online, November 16-20, 2020, pages 2339–2349. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.183. URL https://doi.org/10.18653/v1/2020.emnlp-main.183.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/ARXIV.2306.13549. URL https://doi.org/10.48550/arXiv.2306.13549.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 381–388. AAAI Press, 2015. doi: 10.1609/AAAI.V29I1.9179. URL https://doi.org/10.1609/aaai.v29i1.9179.

Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In Erik Cambria, Bing Liu, Yongzheng Zhang, and Yunqing Xia, editors, *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013*, pages 10:1–10:8. ACM, 2013. doi: 10.1145/2502069.2502079. URL https://doi.org/10.1145/2502069.2502079.