MET-Meme: a Multimodal Meme Dataset Rich in Metaphors

Bo Xu*
School of Software, &Key Laboratory for Ubiquitous Network and Service Software of Liaoning
Dalian University of Technology
Dalian, Liaoning, China
boxu@dlut.edu.cn

Tingting Li
School of Software, &Key Laboratory for Ubiquitous Network and Service Software of Liaoning
Dalian University of Technology
Dalian, Liaoning, China
22017006@mail.dlut.edu.cn

Junzhe Zheng
School of Software, &Key Laboratory for Ubiquitous Network and Service Software of Liaoning
Dalian University of Technology
Dalian, Liaoning, China
1360281502@mail.dlut.edu.cn

Mehdi Naseriparsa
Global Professional School
Federation University Australia
Ballarat, Victoria, Australia
m.naseriparsa@federation.edu.au

Zhehuan Zhao
School of Software, &Key Laboratory for Ubiquitous Network and Service Software of Liaoning
Dalian University of Technology
Dalian, Liaoning, China
z.zhao@dlut.edu.cn

Hongfei Lin
School of Computer Science and Technology
Dalian University of Technology
Dalian, Liaoning, China
hflin@dlut.edu.cn

Feng Xia
School of Engineering, IT, and Physical Sciences
Federation University Australia
Ballarat, Victoria, Australia
f.xia@ieee.org

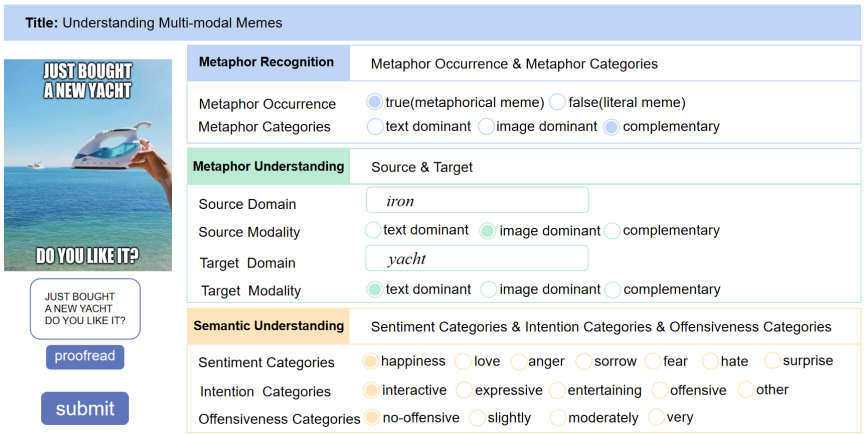


Figure 1: The Annotation Template of Multi-modal Memes in MET-Meme

ABSTRACT

Memes have become the popular means of communication for Internet users worldwide. Understanding the Internet meme is one of the most tricky challenges in natural language processing (NLP) tasks due to its convenient non-standard writing and network vocabulary. Recently, many linguists suggested that memes contain rich metaphorical information. However, the existing researches ignore this key feature. Therefore, to incorporate informative metaphors into the meme analysis, we introduce a novel multimodal meme dataset called *MET-Meme*, which is rich in metaphorical features. It contains 10045 text-image pairs, with manual annotations of

*corresponding author

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532019

the metaphor occurrence, sentiment categories, intentions, and offensiveness degree. Moreover, we propose a range of strong baselines to demonstrate the importance of combining metaphorical features for meme sentiment analysis and semantic understanding tasks, respectively. *MET-Meme*, and its code are released publicly for research in https://github.com/liaolianfoka/MET-Meme-A-Multimodal-Meme-Dataset-Rich-in-Metaphors.

CCS CONCEPTS

• **Computing methodologies** → *Language resources*; *Lexical semantics*.

KEYWORDS

meme dataset; multimodal learning; metaphor; sentiment analysis

ACM Reference Format:
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), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3477495.3532019

1 INTRODUCTION

The Internet meme is defined as a cultural influence that is widely spread and replicated among users via the Internet, primarily for humor [4]. Social media platforms such as Facebook and Twitter provide the most fertile soil for the reproduction and spread of these memes [7]. There is a strong correlation between emotions in social media and the memes that are produced by the users [9]. Therefore, meme interpretation plays a pivotal role in effectively capturing and analyzing the genuine semantic expression of the users. Furthermore, it is imperative to correctly capture the semantics of memes, which facilitate many NLP tasks such as question and answer, sentiment analysis, and intention detection.

Recently, memes' high prevalence has attracted the attention of linguists. These linguists have illustrated that memes always contain inherent metaphorical characteristics which occur in a multimodal way ([1], [19]). For example, in Figure 2(a), the metaphorical message of *"lose your phone"* is conveyed by a mapping between the target domain *"phone"* and the source domain *"oxygen"* from two modalities. This meme infers that the phone is as essential as the oxygen for people. Figure 2(b) offers another example with the metaphor of *"head made from monitor"*; thus, a relation is triggered between the two different entities, *head(job)* and *monitor*, with the perceptual idea that some people constantly monitor others. The source domain *"monitor"* comes from the image, while the target domain *"head(job)"* appears in both text and image. To deeply understand the multimodal memes, it is a requirement to decode the metaphorical messages. Therefore, the decoding process is an integral part of the meme study.

However, current meme studies mainly focus on multimodal learning. They explored the correlation between text and images for sentiment analysis tasks ([9], [12]). The correlation is captured by adopting deep neural networks with image and text features; then, they classify the memes into different types of emotions ([2], [12]). In addition, a few multimodal datasets have been created for meme studies in recent years. Sharma *et al.*[24] released the meme dataset called *MEMOTION* for sentiment analysis, which is a big step forward for accomplishing the meme interpretation task. "Hate meme" datasets for different languages are also generated for hate detection ([15], [13]). However, the metaphorical characteristics in memes have not received the full attention they deserve [35]. That's partly due to the severe lack of multimodal metaphor meme datasets, which are challenging, time-consuming, and labor-intensive to create.

(a) when you lose your phone for a week and you find it; Phone

(b)Someone's job in lives

Figure 2: Examples of Metaphorical Memes

To overcome the above limitations, in this paper, we constructed a multimodal metaphor meme dataset called *MET-Meme* which spans across two languages, amounting to 10045 text-image pairs with manual annotations. *MET-Meme* expands meme understanding with metaphor characteristics and improves the performance of automatic semantic and sentiment comprehension by investigating meme's metaphor cues (Figure 1). Our main contributions are as follows:

(1) We generate a novel multimodal metaphor meme dataset, namely *MET-Meme* (10045 text-image pairs). It will be released publicly to facilitate NLP studies. We present the fine-grained manual annotations of the metaphor occurrence, sentiment categories, intention, and offensiveness, respectively. Moreover, we thoroughly describe the quality control process and agreement analysis for multiple annotators.
(2) To our knowledge, we are the first to consider metaphor features for meme studies. Our experimental results quantitatively verify the role of metaphor features. Furthermore, the results demonstrate to what extent the metaphor affects the distribution of sentiment, intention, and offensiveness.
(3) We propose four tasks to evaluate fine-grained multimodal memes understanding, including metaphor detection, sentiment analysis, intention detection, and offensiveness detection. A range of baselines with benchmark results are reported to verify the potential and fruitfulness of *MET-Meme* for future research.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 formally presents the multi-modal meme dataset *MET-Meme* including memes collection, data annotation and annotation consistency. Section 4 analyzes the distribution of *MET-Meme*. Section 5 presents our baseline models. Section 6 describes experimental settings and metrics. Besides, Section 6 also analyzes the results to illustrate the effectiveness of metaphor. Finally, Section 7 concludes the paper.

2 RELATED WORK

The related work of this paper focuses on the research of memes. It is divided into two parts: (a) meme datasets, and (b) meme semantic understanding.

2.1 Meme Dataset

A few multimodal datasets have been created for meme studies in recent years. Sharma *et al.*[24] released a sentiment analysis challenge based on the meme dataset called *MEMOTION* (10k text-image pairs), which is a big step forward for the task of meme's understanding. Meanwhile, the meme, a language that is virally spread on the Internet, raises the opportunity for hate speech to lurch. The hatred, satire, and offense detection in memes have attracted the attention of many scholars. Different meme datasets have been released for detecting negative information or sentiments in the NLP community, including several popular ones, e.g., a multimodal misogyny dataset (800 text-image pairs) [10], a multimodal hateful meme dataset (over 10k text-image pairs) [13], and a multimodal wild hateful meme dataset (2840 text-image pairs) [15]. The Hateful Memes Challenge[13] is the first competition that is created to focus on detecting multimodal meme hate speech.

Although many scholars have constructed multimodal meme datasets, they only labeled sentiment categories. Therefore, the metaphorical characteristics have not received the full attention they deserve. However, some linguists have suggested that the cognitive nature of memes implies that the integrated metaphor information may contribute to better meme understanding([1], [19]). To incorporate the rich metaphor information, we introduce our dataset *MET-Meme*, which is large scale and contains both sentiment and metaphor annotations. *MET-Meme* is different from existing datasets and is effective for NLP studies.

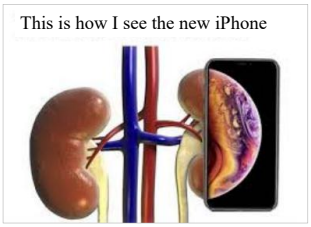


Figure 3: An Annotation Example of Metaphorical Memes

2.2 Meme Understanding

Sentiment analysis, the main research field in the NLP community, has produced many datasets in the text domain; however, meme sentiment analysis is rarely considered. Some scholars explored the correlation between text and images for sentiment analysis tasks and found meme can contribute to sentiment analysis in social media ([9], [12]). With the release of multimodal meme datasets, some scholars adopted deep neural networks with image and text features for sentiment classification on *MEMOTIN* ([2], [22]). To detect the presence of offensive content, some scholars utilized different visual and linguistic transformer models to make a breakthrough. ([31], [37]).

The existing methods only took text and image features into account while they ignored the vitally important metaphorical features. It is partly due to the severe lack of multimodal metaphor meme datasets with their challenging, time and labor-consuming creation. Thus, our *MET-Meme* opens the door to automatic meme understanding by investigating metaphor cues.

3 THE MET-MEME DATASET

3.1 Dataset Collection

To create a large-scale multimodal metaphor meme dataset to support the research on better understanding memes, we collect publicly available data from a range of sources, including social media (Twitter and Weibo), Google, and Baidu images. We did not store any personal data such as user IDs and usernames to protect the users' privacy. Table 1 presents an overview of the statistics that are captured from the dataset. We consider meme types as keywords to download memes from Google, Baidu, and Weibo images.

For the Chinese meme dataset, we identified a total of six unique categories, including animals, scenery, animations, films, dolls, and humans. We collected a total of 12,369 jpg format memes. Meanwhile, the English meme dataset (4000 text-image pairs) consists of two parts, one from *MEMOTION* and the other from Google search. Then, we removed the non-memes and the memes whose both text and images are repeated. All the memes must contain a clear background picture and textual content. So far, this data has accumulated 10045 text-image pairs, including 6045 Chinese and 4000 English memes, respectively.

Table 1: MET-Meme Dataset Statistics

Item(Samples)	Bilingual	Chinese	English
Total	10045	6045	4000
Metaphorical	3441(34%)	2327(39%)	1114(28%)
Literal	6604(66%)	3718(61%)	2886(72%)
Total Words	90191	42844	47347
Average Words	9	7	12

3.2 Metaphor Annotation

Our annotations focus on whether the meme has a metaphorical expression or not; thus, the memes are either annotated by metaphorical or literal labels. Figure 3 contains complete metaphor annotation with an example meme. The metaphorical label indicates that there is a metaphorical occurrence in this meme, while the literal label indicates the opposite. Furthermore, we divide metaphorical memes into three sub-categories: text dominant, image dominant, and complementary [28]. The three sub-categories are illustrated in Figure 4 (page 4). Text dominant (Figure 4(a)) means that the text itself is sufficient to convey metaphorical information. For example, *"you are my sunshine"* can be identified as metaphorical expressions only by the text content. By contrast, in the image dominant sub-category, images play the dominant role in conveying metaphorical information and they provide sufficient information for readers to understand the metaphors. In Figure 4(b), we observe the metaphorical message *"ear is a snake(toxic)"* from the

image. The complementary category involves a roughly equal role of texts and images in rendering metaphorical information. If texts and images are interpreted separately, metaphors cannot be understood. In Figure 4(c), when people read the text, *"I occasionally enjoy cocaine."*, they do not realize the metaphorical use until they observe the snow in the corresponding image and infer that the target *"snow"* is expressed in terms of the source *"cocaine"*.

Figure 4: Examples of Metaphor

3.3 Sentiment Annotation

Sentiment analysis, as one of the most important topics in NLP, has attracted a large number of scholars([38], [23], [3], [34]). We followed the sentiment classification method that is proposed in the "Sentiment Vocabulary Ontology" [33]. This method has set up seven sentiment categories as follows: happiness, love, anger, sorrow, fear, hate, and surprise. Each category has been fine-tuned to fit memes' specific emotional style:

(1) **happiness:** positive sentiments of happiness, relaxation, peace, striving to be strong;
(2) **love:** positive sentiments of desire to explore, yearning, liking, worship, respect, and approval;
(3) **anger:** express negative feelings of anger or irritation;
(4) **sorrow:** including but not limited to sadness, disappointment, grievance, guilt, stress, and despair;
(5) **fear:** negative emotions of panic, fear, avoidance, and apology;
(6) **hate:** negative sentiments of boredom, hatred, demeaning, jealousy, and disgust;
(7) **surprise:** express disbelief, surprise and feelings when things are not expected to happen.

Figure 3 describes sentiment annotation with an example meme. This meme is labelled as *"sorrow"*.

3.4 Intention Annotation

The memes' intentions have been widely studied by scholars. We summarize these intentions as follows: expression, entertainment, social integration, etc ([17], [29], [6]). Inspired by eight intentions on tweets [16], we design intention categories by combining the characteristics of memes. Since a meme may contain multiple purposes, we only consider the most important purpose:

(1) **interactive:** highlight closeness and interaction between the two parties, hoping to get or reduce feedback;
(2) **expressive:** express feelings, create own public persona;
(3) **purely entertaining:** without other intentions, just for the purpose of being funny;
(4) **offensive:** discriminates, satirize, and abuses others' occupation, gender, appearance or insults others' personality;
(5) **other:** do not meet any of the above four intents.

Figure 3 presents a specific example of intention annotation. This meme is labelled as *"offensive"*.

3.5 Offensiveness Annotation

With the growing volume of multi-modality in social media, the prevalence of meme's malicious speech, including hate, racism, and misogyny, has become a nightmare for many social media companies as well as scholars ([15], [18]). The detection of offensive content is an ongoing struggle, which encouraged us to propose this task. Our dataset is applicable for fine-grained classification tasks. From the ethics and law perspectives, we aim to detect the main content that is not suitable for public dissemination:

(1) **label–0:** non-offensive; no intention to attack, satirize, disparage or offend;
(2) **label–1:** slightly offensive; may cause slight discomfort to others, not excessively;
(3) **label–2:** moderately offensive; destructive, should not be used in normal communication;
(4) **label–3:** very offensive; not suitable to appear on any public platform, causing harm to others' feelings and reputation.

As shown in Figure 3, the example is a *"moderately offensive(2)"* meme.

3.6 Annotation Process

We adopted two independent annotation approaches for accomplishing different tasks, including conceptual metaphor understanding, sentiment, intent, and offensiveness annotations, respectively. We describe the components and workflow of the full annotation pipeline shown in Figure 5 (page 5) below.

3.6.1 Metaphor Annotation Process. Since the conceptual metaphor understanding requires relevant professional knowledge, we recruited expert annotators to perform the daunting annotation task. The metaphor annotation focuses on the metaphorical relationship between source and target domains, respectively. By considering the previous experience of understanding metaphor [30], we adopted the methods of "adjective-noun" and "verb-noun" to annotate source and target domains, respectively. The annotator team was comprised of 12 NLP postgraduate students and 8 research assistants who were fully familiar with metaphor. They performed metaphor occurrence, metaphor category, and domain category annotations, respectively. Each meme was annotated by at least three graduate students and two research assistants.

3.6.2 Sentiment, Intention, Offensiveness Annotation Process. We handed over the semantic annotation tasks to a professional crowdsourcing company. These tasks include sentiment, intention, and offensiveness labeling. All annotators were presented with both the text and images randomly. The memes' text was obtained by OCR APIs [26]. The annotator had to manually proofread it. Each meme was annotated by at least 3 annotators. Moreover, all the

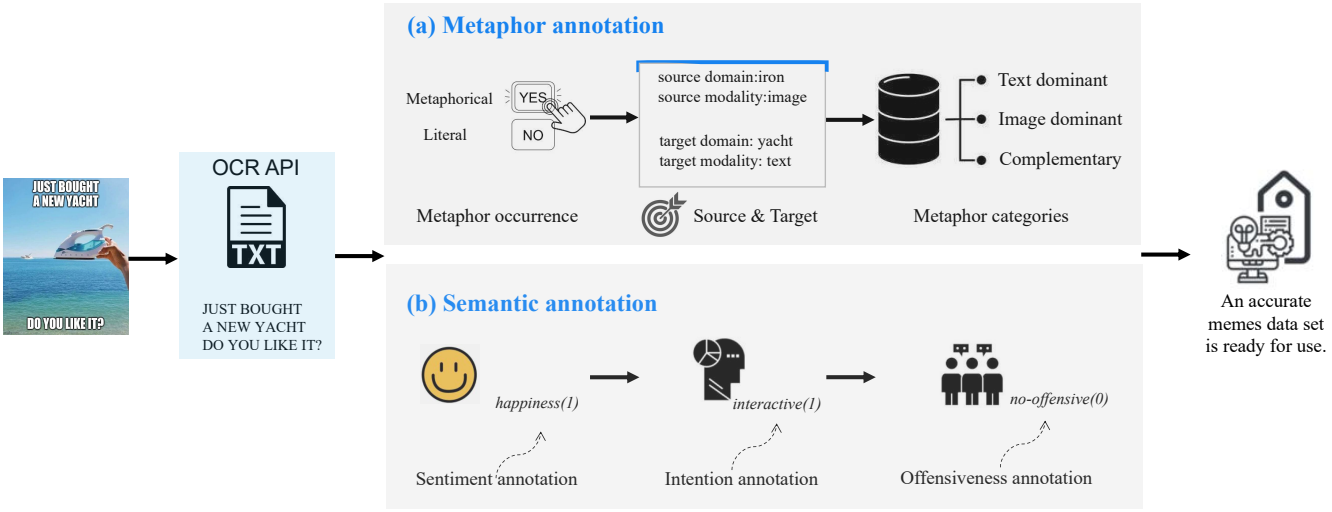


Figure 5: The Annotation Pipeline with an Example Meme.

annotators could select sentiment types, intent, and offensiveness with detailed annotation instructions.

3.6.3 Cost. We adopted two independent approaches, including expert-based, and crowd-sourcing-based for performing the annotation process. We performed different tasks such as metaphor and semantic annotations in the process. Though manual labeling is time-consuming and costly, we reached an agreement with the crowd-sourcing company on compensation and payment methods. To ensure the data quality and a fair salary for labeling personnel, we paid $0.5 for each qualified meme labeling and paid off all labor wages immediately after they finished labeling.

3.7 Quality Control and Annotation Consistency

We established strict criteria for the annotation process to minimize the subjectivity of the annotators. For example, all the annotators have attended metaphor guideline meeting to discuss the labeling of images and other details. Furthermore, the annotation process meeting should be held once a week for more than 2 hours to discuss annotation problems and matters that need attention. Each metaphorical meme must be approved by at least three annotators. If any annotator strongly disagrees with the annotation result, the meme will be discarded.

We introduced strict labeling rules with the crowd-sourcing company to perform the sentiment, intention, and offensiveness annotation tasks. For example, each meme should be annotated by at least three annotators. The annotators need to observe and understand the meme 30 seconds before the annotation process. Simultaneously, we continue to follow the entire annotation process and communicate with the person in charge of labeling every $2-3$ days to ensure the rationality of data labeling. Furthermore, we remove the workers who consistently produce low-quality work.

To perform the task, we adopted kappa score, k, in Fleiss kappa tests [8] to measure the inter-annotator agreements. As shown in Table 2, the agreement on sentiment category identification was $k = 0.815$; the intention category identification was $k = 0.817$; the offensiveness detection identification was $k = 0.825$, which means they are substantially reliable.

Table 2: Fleiss Kappa Score(k) on Different Annotation Tasks

	sentiment	intention	offensiveness
k	0.815	0.817	0.825

4 DATASET ANALYSIS

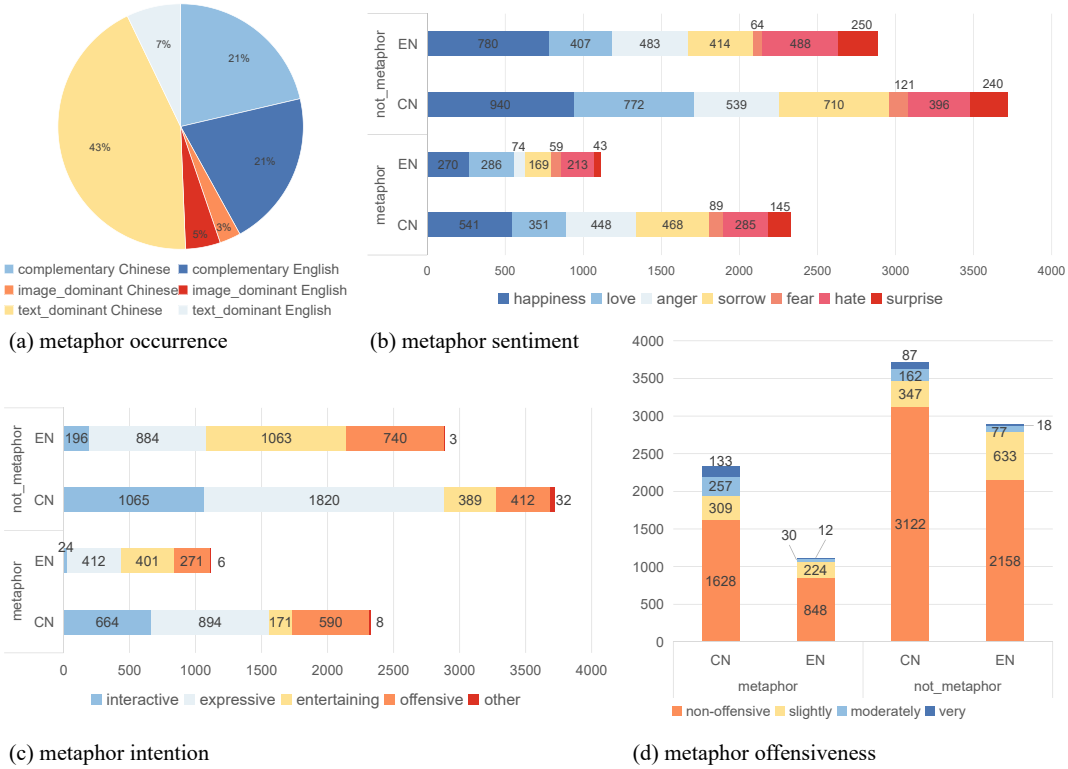
4.1 Metaphor Occurrence

Table 1 presents an overview of the *MET-Meme* dataset statistics. The number of Metaphorical samples represents the number of memes that contain metaphor's expression. By contrast, Literal samples represent the memes without the occurrence of metaphors. From Table 1, we observe that 34% of memes contain metaphorical features.

In addition, we analyzed the role of text and images in understanding metaphors. Figure 6(a) (page 6) represents the distribution of all metaphor categories, including text dominant, image dominant, and complementary. We observe that a text dominant category accounts for the largest proportion of metaphors which is 50%. That's because texts have rich grammatical features and diverse forms. The next highest proportion is the complementary category. Its proportion is almost 42%. This verifies that the interplay of textual content and visual modality plays an instrumental role to understand metaphor meaning within the metaphors' occurrence.

4.2 The Distribution of Sentiment Categories

Figure 6(b) represents metaphor sentiment distribution. In the literal memes, negative sentiments accounted for 48.69%, including anger (15.48%), sorrow (17.02%), fear (2.80%), and hate (13.39%). In



(a) metaphor occurrence (b) metaphor sentiment

(c) metaphor intention (d) metaphor offensiveness

Figure 6: Distribution of MET-Meme Dataset (CN represents Chinese memes and EN represents English memes)

the metaphorical memes, negative sentiment accounted for 52.46%, including anger (15.17%), sorrow (18.51%), fear (4.30%), and hate (14.48%). It turns out that there are more negative sentiments conveyed by metaphorical expressions.

4.3 The Distribution of Intention Categories

As depicted in Figure 6(c), expressive intentions occur most frequently in the English metaphorical memes, which are accounted for 36.98%, while the most widely used literal memes are for entertainment purposes, which are accounted for 36.83%. Therefore, English-speaking people are more likely to use metaphorical expressions when expressing feelings. This conclusion is consistent with [36], which shows that people are more accustomed to using metaphors when expressing emotions, whether in tweet, advertisements or memes.

4.4 The Distribution of Offensiveness Categories

Figure 6(d) demonstrates that there are differences in the offensiveness distribution between the metaphorical and literal memes. The most obvious difference occurred in the offensive intent of Chinese memes, where 16.03% of literal memes are offensive, while in the metaphorical memes, offensive memes accounted for 30.04%. Our findings accord with the results of previous findings in hate meme studies, that the metaphors convey more implicitly offensive messages than the texts [13].

From each picture in Figure 6, metaphorical features cannot be ignored. Next, we will prove this through experiments.

5 BASELINE MODELS

We propose four tasks and corresponding baselines for our released dataset. These tasks are as follows: metaphor recognition, multimodal sentiment analysis, multimodal intention detection, and multimodal offensiveness detection. The general outline of the architectures of our model for the four tasks is presented in Figure 7 (page 7). We describe each model component below.

5.1 Encoders

While the end classifier or decoder for each task is different, we use the same set of encoders based on text embedding, image embedding, metaphor embedding, and common neural network architecture.

5.1.1 Multilingual Bert. The text and metaphor inputs are processed similarly. We adopt Google's Bert_multi_cased pre-training model, and Multilingual Bert ([5], [32]). Multilingual Bert employs the WordPiece algorithm for tokenization where the dictionary size is 110K. We take the mean value of the penultimate hidden state value on the token for the text representation. The same process applies to metaphor representation. We denote the text feature

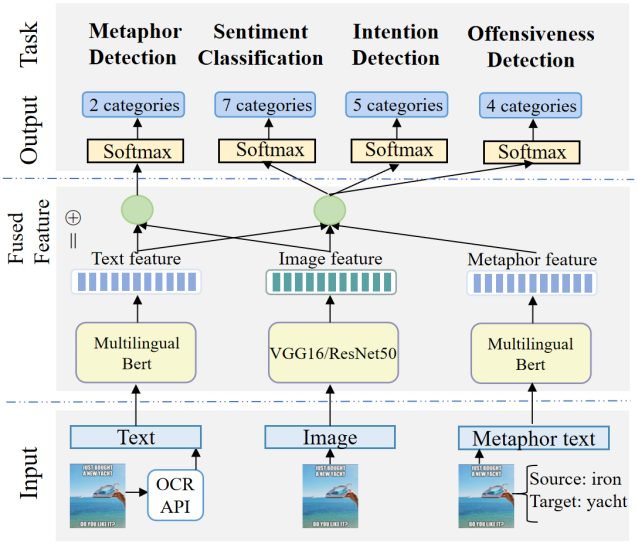


Figure 7: General Model Architectures for Four Different Tasks

set as Te, with the embedding space belonging to R^{N*768} and the metaphorical feature sets are S(Source domain feature set) and Ta(Target domain feature set), respectively, with the embedding space belonging to R^{N*768}. Among them, N represents the number of memes.

5.1.2 VGG16. We implement the pre-trained convolutional neural network-based classifiers VGG16 to process the image [25]. It has 5 max-pooling layers and 16 weighted layers. We take a meme with a size of 224×224×3 as input and use the output of the second-last layer for the image representation. We denote the image representation set as V, and the embedding space belongs to R^{N*4096}.

5.1.3 ResNet50. We also adopt ResNet50, a classic neural network for computer vision tasks, to obtain representations of meme images [11]. It has a 7×7×64 convolutional layer, 16 3-layer building blocks, and a fully connected layer. We use the output of the second-last layer for the image representation. We denote the image representation set as R while the embedding space belongs to R^{N*2048}.

5.2 Feature Fusion

For feature fusion, we reduce the dimension of feature vectors in image feature sets V and R from 4096 and 2048 dimensions to 768 dimensions. For input meme i, we get the text representation Te_i, the source representation S_i, the target representation Ta_i, the VGG image representation V_i, and the ResNet50 image representation R_i.

5.2.1 Concatenation. We simply splice the different feature vectors together through mainstream early fusion method concatenation [27]. The fused vectors $f1_i$ is denoted as:

$$f1_i = [Te_i, V_i(orR_i), S_i, Ta_i] \tag{1}$$

5.2.2 Element-wise Add. For input meme i, we firstly spliced text, source and target into "[CLS] text [SEP] source [SEP] target [SEP]" structure, which is taken as the input of Multilingual BERT. The corresponding 768 - dimensional fusion feature T_i is obtained. Then, we also use another popular early fusion method element-wise add to fuse feature vector T_i and image feature V_i (or R_i) [20]. We add feature vectors element by element, and take the average vector as the final feature vector $f2_i$:

$$f2_i = \frac{\sum(T_i, V_i(orR_i))}{2} \tag{2}$$

5.3 Tasks

5.3.1 Metaphor Understanding. The model is given with the text content and image features. Finally, we input the fused feature vectors into the full connection layer and adopt the softmax activation function to produce the prediction results of two categories: literal meme, and metaphorical meme.

5.3.2 Sentiment Analysis. The sentiment analysis is a seven-category classification task. The inputs to our model are as follows: meme's text content, visual features, and metaphorical information. After we get the fused feature vector, we input it into the full connection layer. Then, we adopt the softmax activation function to get the prediction results where we have seven categories: happiness, love, anger, sorrow, fear, hate, and surprise.

5.3.3 Intention Detection. The intention detection is defined as a five-category classification task. The inputs include meme's text, visual features, and metaphorical information. We use the same architecture from Section 5.3.2 to build the classifier for predicting results. The five categories for the prediction task are as follows: interactive, expressive, purely entertaining, offensive, and other.

5.3.4 Offensiveness Detection. The offensiveness detection is defined as a four-category classification task. We input meme's text content, visual features, and metaphorical information and adopt the same classifier architecture in Section 5.3.2 to predict the following labels: non-offensive, slightly offensive, moderately offensive, and very offensive.

6 EXPERIMENTS

Our experiments are mainly divided into four tasks: metaphor understanding, sentiment analysis, intention detection, and offensiveness detection. We also explore whether the metaphor feature is beneficial to performance improvement. Furthermore, our results investigate the effects of features from different modes and multi-modal fusion methods on performance.

6.1 Preprocessing and Experiment Settings

For each meme's text, we removed external links, non-Chinese or non-English texts, and strange symbols such as emojis, URLs, and numbers. Furthermore, all letters are converted to lowercase. The metaphor features of metaphorical memes are the source and target domains, while that of literal memes are uniformly represented by "[CLS] [SEP]". To avoid problems caused by different image

Table 3: F1-value Results of Different Tasks on MET-Meme

Metaphor Understanding

				Chinese-English		English		Chinese	
	text	image	fusion	val	test	val	test	val	test
random	–	–	–	0.5339	0.5339	0.4971	0.5109	0.4676	0.4834
text	Multilingual Bert	–	–	0.7534	0.7353	0.7565	0.7550	0.7624	0.7718
image	–	Vgg16	–	0.6493	0.6526	0.7474	0.7741	0.6087	0.6304
	–	Resnet50	–	0.6555	0.6389	0.7270	0.7411	0.6000	0.6535
Text + Image	Multilingual Bert	Vgg16	add	0.7501	0.7490	0.7835	0.8133	0.7447	0.7404
	Multilingual Bert	Resnet50	add	0.7371	0.7249	0.7909	0.7931	0.7562	0.7718
	Multilingual Bert	Vgg16	cat	0.7558	**0.7602**	0.7977	**0.8239**	0.7652	0.7290
	Multilingual Bert	Resnet50	cat	0.7380	0.7303	0.7839	0.7999	0.7701	**0.7723**

Sentiment Analysis

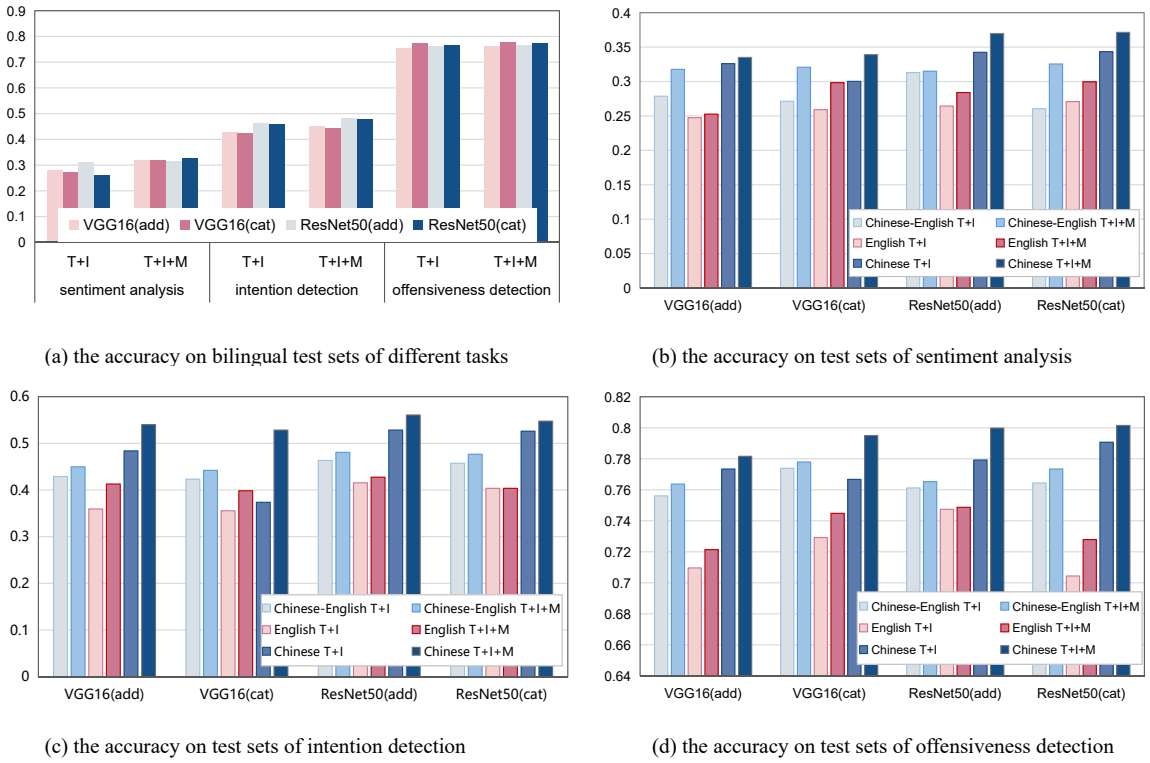
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.1533	0.1621	0.1252	0.1340	0.1601	0.1683
Multilingual Bert	Vgg16	–	add	0.2763	0.2586	0.2431	0.2363	0.3103	0.3096
Multilingual Bert	Vgg16	Multilingual Bert	add	0.2840	0.2935	0.2911	0.2465	0.3442	0.3250
Multilingual Bert	Vgg16	–	cat	0.2599	0.2507	0.2465	0.2516	0.3043	0.2836
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.3007	0.2846	0.2329	**0.2768**	0.3542	0.3342
Multilingual Bert	Resnet50	–	add	0.2980	0.2922	0.2506	0.2566	0.3367	0.3297
Multilingual Bert	Resnet50	Multilingual Bert	add	0.3197	**0.3071**	0.2708	0.2739	0.3782	0.3557
Multilingual Bert	Resnet50	–	cat	0.2874	0.2549	0.2506	0.2487	0.3196	0.3345
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.2947	0.2761	0.2526	0.2507	0.3881	**0.3690**

Intention Detection

				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.2246	0.2320	0.2320	0.2232	0.2232	0.2278
Multilingual Bert	Vgg16	–	add	0.4493	0.4269	0.3923	0.3491	0.4987	0.4737
Multilingual Bert	Vgg16	Multilingual Bert	add	0.4689	0.4492	0.4006	0.4032	0.5167	0.5293
Multilingual Bert	Vgg16	–	cat	0.3993	0.4163	0.3797	0.3461	0.3971	0.3612
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.4551	0.4390	0.3979	0.3856	0.4914	0.5158
Multilingual Bert	Resnet50	–	add	0.4840	0.4596	0.3717	0.4016	0.5361	0.5207
Multilingual Bert	Resnet50	Multilingual Bert	add	0.5025	**0.4800**	0.3764	**0.4165**	0.5474	**0.5498**
Multilingual Bert	Resnet50	–	cat	0.4867	0.4480	0.3775	0.3836	0.5167	0.5098
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.4997	0.4549	0.4231	0.3922	0.5509	0.5408

Offensiveness Detection

				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.3395	0.3192	0.3089	0.3659	0.3218	0.3283
Multilingual Bert	Vgg16	–	add	0.6548	0.7054	0.6463	0.6652	0.6902	0.7384
Multilingual Bert	Vgg16	Multilingual Bert	add	0.7081	0.7154	0.6790	**0.6839**	0.7060	0.7601
Multilingual Bert	Vgg16	–	cat	0.6507	0.6763	0.6405	0.6394	0.6808	0.7312
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.6698	0.6929	0.6418	0.6725	0.6984	0.7319
Multilingual Bert	Resnet50	–	add	0.6635	0.7131	0.6427	0.6418	0.7017	0.7514
Multilingual Bert	Resnet50	Multilingual Bert	add	0.7031	**0.7205**	0.6653	0.6549	0.7287	**0.7664**
Multilingual Bert	Resnet50	–	cat	0.6570	0.6729	0.6444	0.6600	0.6805	0.7512
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.6579	0.6801	0.6642	0.6746	0.7006	0.7557



(a) the accuracy on bilingual test sets of different tasks

(b) the accuracy on test sets of sentiment analysis

(c) the accuracy on test sets of intention detection

(d) the accuracy on test sets of offensiveness detection

Figure 8: The Accuracies on Test Sets of Different Tasks

Table 4: Hyperparameters

Hyper-Parameter	Value
Dropout	0.4
Batch size	48
Learning rate	1e-5

sizes, images are sent into classifiers to be adjusted to $224 * 224 * 3$. Meanwhile, we randomly divided our dataset into training/val/test sets by using 60/20/20% split.

We adopt Pytorch to build our model [21]. All training tasks use the same parameters, as presented in Table 3. When the model tends to converge, we update the parameters of the pre-training model with training data to avoid the over-fitting issue. In addition, we adopt Adam optimizer to optimize the loss function [14].

6.2 Metrics

To assess the classification task performance, we report accuracy and weighted F1-value, Precision, Recall as measurement indicators. To improve readability, we report F1-value in page 8, denoted by F1 in Equation 3, and accuracy in Figure 8, for metaphor recognition, multimodal sentiment analysis, multimodal intention detection, and multimodal offensiveness detection tasks. Furthermore, the experimental results of Precision (Table 5) and Recall (Table 6) are available in the appendix (page 12, and page 13).

We denote L as the number of categories, w_i as the weight (the proportion of samples) for the ith category, TP as true positives, TN as true negatives, FN as false negatives and FP as false positives.

$$F1 = \frac{\left(2 * Precision_{weighted} * Recall_{weighted}\right)}{\left(Precision_{weighted} + Recall_{weighted}\right)} \tag{3}$$

$$Precision_i = \frac{(TP_i)}{(TP_i + FP_i)} \tag{4}$$

$$Precision_{weighted} = \frac{\sum_{i=1}^{L}(Precision_i * w_i)}{|L|} \tag{5}$$

$$Recall_i = \frac{(TP_i)}{(TP_i + FN_i)} \tag{6}$$

$$Recall_{weighted} = \frac{\sum_{i=1}^{L}(Recall_i * w_i)}{|L|} \tag{7}$$

$$accuracy = \frac{TP_i + TN_i}{TP_i + TN_i + FN_i + FP_i} \tag{8}$$

6.3 Results

We present the performance results on the test set (test) and the validation set (val). Tables 4 (page 8) presents the results for the following tasks: metaphor detection, sentiment analysis, intention detection, and offensiveness detection. Figure 8(page 9) presents the accuracy results on bilingual test sets of sentiment analysis, intention detection, and offensiveness detection, where "T+I" means only text and image features, while "T+I+M"means features include text, image and metaphorical features.

6.3.1 Metaphor Understanding. As presented in Table 4, our experiments on metaphor understanding are divided into four parts: random, only text, only image, and multimodal models. We utilized two different visual pre-training models called VGG16 and Resnet50 in the multimodal models part. Also, we adopted two fusion methods: element-wise add (add) and concatenation (cat).

No matter in which dataset, the best results of metaphor understanding occurs in the multimodal models. Models with "Mulitilingual Bert + VGG16 + cat" can achieve the best performance with 76.02% and 82.39% on bilingual and English test set. Meanwhile, on Chinese test set, "Mulitilingual Bert + Resnet50 + add" achieves best performance with 77.18%. On the whole, although text plays an important role in metaphor understanding, simple multimodal fusion can still achieve good results. Furthermore, we observe that the largest improvement for multimodal fusion occurs in English memes, which achieves a 4.98% improvement on the test set.

6.3.2 Sentiment Analysis. Multimodal models are used in the sentiment analysis tasks (Multilingual Bert for text and metaphor, VGG16 and Resnet50 for images, with element-wise add (add) and concatenation (cat) feature fusion methods).

Table 4 summarizes the results of sentiment analysis. "Multilingual Bert + Resnet50 + add" achieves the best performance with 30.71% on the bilingual test set. "Multilingual Bert + Resnet50 + cat" achieves the best performance with 36.90% on the Chinese test set. The performance of the sentiment analysis task is quite lower compared with other semantic understanding tasks. That's because it is a seven classification task. There is still little room for improvement in the model. But as expected, we observe a moderate improvement in the F1-value when we adopt more compositional input vectors (text, image, and metaphor). Take the Chinese test set as an example; after incorporating metaphorical features, the "Multilingual Bert + VGG16 + cat" model performance improved by 5.06%.

Similarly, Figure 8(b) shows that integrating metaphorical features is helpful to improve the accuracy of sentiment analysis task. In bilingual dataset, "Multilingual Bert + Resnet50 + cat" has the best improvement performance, with a 6.5% improvement.

6.3.3 Intention Detection. We add metaphor features into input feature vectors to explore the role of metaphor features in the intention detection task. Our intention detection task has different multimodal models (Multilingual Bert for text and metaphor, VGG16 and Resnet50 for images), fusion methods (element-wise add and concatenation), and input features (text, image, and metaphor).

From Table 4, the best performance for intention detection on all test sets are achieved by "Multilingual Bert + Resnet50 + add"where the results stand at 48.00%, 41.65%, and 54.98%.

From Table 4 and Figure 8(c), significant performance improvements can be found across almost all data sets after fusing metaphorical features, confirming that simple multimodal fusion can help the model learn more. On the Chinese dataset, the concatenation of Multilingual Bert and VGG16 increased F1-value by 15.46% and accuracy by 15.48%on the test set, which indicates metaphor plays an important role.

6.3.4 Offensiveness Detection. We design different multimodal models (Multilingual Bert for text and metaphor, VGG16 and Resnet50 for images), fusion methods (element-wise add and concatenation), and input features (text, image, and metaphor) to explore whether metaphorical features contribute to the offensiveness detection task.

From the semantic experimental results in Table 4, our model performs best on offensiveness detection tasks. The highest F1-value of the model reaches 76.64% on the Chinese test set when "Multilingual Bert + Resnet50 + add" is applied. With the addition of metaphorical features, the most obvious performance improvement is in "Multilingual Bert + VGG16 + cat" on the English test set, which improved by 3.31%.

According to the accuracy results in Figure 8(d), "Multilingual Bert + VGG16 + cat" performs best in the bilingual test set, scoring 77.79%. With the addition of metaphor feature, the best performance improvement occurred in "Mulitilbert + VGG16 + cat" on the Chinese test set, improving by 2.82%.

No matter the accuracy of all tasks in Figure 8 (a) or all F1-values in Table 4, the best performances in the above sentiment analysis and semantic understanding tasks all occur when the text, image, and metaphor features are all supplied to the input. The observed performance boost, which is facilitated by adding metaphor features, indicates that the models learn more even by fusing metaphor features with the most basic methods. This result verifies that metaphor feature fusion is fruitful to classify NLP tasks at the current time step. Hence, we believe that the performance will be significantly improved in the future work by optimizing the model and integrating metaphors.

7 CONCLUSION

This paper presents the creation of a novel resource, a large-scale multimodal metaphor meme dataset, *MET-Meme*, with manually fine-grained annotation for meme understanding and research. It also offers a set of baseline results of various tasks and shows the importance of combining metaphorical features for meme sentiment analysis and semantic understanding. *MET-Meme* opens the door to automatic meme understanding by investigating multimodal metaphor cues and their interplay. We hope *MET-Meme* provides future researchers with valuable multimodal metaphor training data for the challenging tasks of meme understanding and sentiment analysis.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62072073, Grant 61906028, Grant 62076046, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT20RC(4)010.

REFERENCES

[1] Stephen Madu Anurudu and Isioma Maureen Obi. 2017. Decoding the Metaphor of Internet Meme: A Study of Satirical Tweets on Black Friday Sales in Nigeria. *AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies* 6, 1 (2017), 91–100.

[2] Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. IIITG-ADBU at SemEval-2020 Task 8: A Multimodal Approach to Detect Offensive, Sarcastic and Humorous Memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 885–890.

[3] Geng Binzong, Yang Min, Yuan Fajie, Wang Shupeng, Ao Xiang, and Xu Ruifeng. 2021. Iterative Network Pruning with Uncertainty Regularization for Lifelong Sentiment Classification. *ACM* (2021).

[4] Patrick Davison. 2012. *The Language of Internet Memes*. 120–134.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Eli Dresner and Susan C Herring. 2010. Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication theory* 20, 3 (2010), 249–268.

[7] Amanda Du Preez and Elanie Lombard. 2014. The role of memes in the construction of Facebook personae. *Communicatio* 40 (09 2014), 253–270. https://doi.org/10.1080/02500167.2014.938671

[8] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[9] Jean H. French. 2017. Image-based memes as sentiment predictors. In *2017 International Conference on Information Society (i-Society)*. 80–85. https://doi.org/10.23919/i-Society.2017.8354676

[10] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2021. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv preprint arXiv:2106.08409* (2021).

[11] 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*. 770–778.

[12] Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. https://doi.org/10.1145/3219819.3219853

[13] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.

[14] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

[15] Hannah Rose Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset. *arXiv preprint arXiv:2107.04313* (2021).

[16] 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. *arXiv preprint arXiv:1904.09073* (2019).

[17] Shao-Kang Lo. 2008. The nonverbal communication functions of emoticons in computer-mediated communication. *Cyberpsychology & behavior* 11, 5 (2008), 595–597.

[18] Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective Dependency Graph for Sarcasm Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1844–1849.

[19] Maria Mitsiaki. 2020. INVESTIGATING METAPHOR IN MODERN GREEK INTERNET MEMES:: AN APPLIED APPROACH WITH L2 PEDAGOGICAL IMPLICATIONS. *Revista Brasileira de Alfabetização* 12 (2020), 73–106.

[20] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. (2019).

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).

[22] Shraman Pramanick, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. Exercise? I thought you said 'Extra Fries': Leveraging Sentence Demarcations and Multi-hop Attention for Meme Affect Analysis. (2021).

[23] Lu Ren, Bo Xu, Hongfei Lin, Xikai Liu, and Liang Yang. 2020. Sarcasm detection with sentiment semantics enhanced multi-level memory network. *Neurocomputing* 401 (2020), 320–326.

[24] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, and Bjorn Gamback. 2020. SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor! (2020).

[25] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

[26] Viswanath Sivakumar, Albert Gordo, and Manohar Paluri. 2018. Rosetta: Understanding text in images and videos with machine learning. *Facebook Engineering blog posted on* 11 (2018), 2018.

[27] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. 32–41.

[28] Miloš Tasić and Dušan Stamenković. 2015. The interplay of words and images in expressing multimodal metaphors in comics. *Procedia-Social and Behavioral Sciences* 212 (2015), 117–122.

[29] Channary Tauch and Eiman Kanjo. 2016. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing: Adjunct*. 1560–1565.

[30] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 248–258.

[31] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. (2020).

[32] Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840* (2019).

[33] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen. 2008. Constructing the affective lexicon ontology. *Journal of the China society for scientific and technical information* 27, 2 (2008), 180–185.

[34] Dongyu Zhang, Minghao Zhang, Teng Guo, Ciyuan Peng, Vidya Saikrishna, and Feng Xia. 2021. In Your Face: Sentiment Analysis of Metaphor with Facial Expressive Features. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[35] Dongyu Zhang, Minghao Zhang, Ciyuan Peng, Jason. J Jung, and Feng Xia. 2021. Metaphor Research in the 21st Century: A Bibliographic Analysis. (2021).

[36] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A Multimodal Dataset for Metaphor Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[37] Xiayu Zhong. 2020. Classification of Multimodal Hate Speech–The Winning Solution of Hateful Memes Challenge. *arXiv preprint arXiv:2012.01002* (2020).

[38] Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, and Hongfei Lin. 2021. Hate Speech Detection Based on Sentiment Knowledge Sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Table 5: Precision Results of Different Tasks on MET-Meme

Metaphor Understanding									
				Chinese-English		English		Chinese	
type	text	image	fusion	val	test	val	test	val	test
random	–	–	–	0.5730	0.5730	0.5584	0.6068	0.4828	0.5084
text	Multilingual Bert	–	–	0.7571	0.7386	0.7565	0.7528	0.7620	0.7713
image	–	Vgg16	–	0.6603	0.6482	0.7513	0.7802	0.6096	0.6283
	–	Resnet50	–	0.6526	0.6345	0.7355	0.7398	0.5972	0.6617
Text + Image	Multilingual Bert	Vgg16	add	0.7509	0.7496	0.7846	0.8149	0.7485	0.7451
	Multilingual Bert	Resnet50	add	0.7367	0.7257	0.7919	0.7909	0.7585	**0.7726**
	Multilingual Bert	Vgg16	cat	0.7583	**0.7597**	0.7969	**0.8269**	0.7652	0.7280
	Multilingual Bert	Resnet50	cat	0.7361	0.7284	0.7846	0.7983	0.7723	0.7718
Sentiment Analysis									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.1803	0.1864	0.1450	0.1541	0.1888	0.1963
Multilingual Bert	Vgg16	–	add	0.2772	0.2540	0.2388	0.2319	0.2984	0.3060
Multilingual Bert	Vgg16	Multilingual Bert	add	0.2821	0.3006	0.2908	0.2452	0.3453	0.3262
Multilingual Bert	Vgg16	–	cat	0.2664	0.2549	0.2447	0.2507	0.3081	0.2830
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.2978	0.2773	0.2926	**0.2841**	0.3515	0.3433
Multilingual Bert	Resnet50	–	add	0.2984	0.2968	0.2506	0.2549	0.3301	0.3428
Multilingual Bert	Resnet50	Multilingual Bert	add	0.3198	**0.3038**	0.2719	0.2724	0.3727	0.3483
Multilingual Bert	Resnet50	–	cat	0.2777	0.2735	0.2452	0.2410	0.3124	0.3730
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.3337	0.2841	0.2579	0.2630	0.3889	**0.3785**
Intention Detection									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.2828	0.2946	0.2946	0.3174	0.3174	0.3134
Multilingual Bert	Vgg16	–	add	0.4597	0.4434	0.4436	0.3548	0.5044	0.4713
Multilingual Bert	Vgg16	Multilingual Bert	add	0.4810	0.4555	0.4112	0.4039	0.5131	0.5268
Multilingual Bert	Vgg16	–	cat	0.4318	0.4275	0.3711	0.3519	0.3893	0.3571
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.4654	0.4518	0.4009	0.3919	0.5172	0.5148
Multilingual Bert	Resnet50	–	add	0.5041	0.4748	0.3661	0.3915	0.5381	0.5149
Multilingual Bert	Resnet50	Multilingual Bert	add	0.5128	**0.4919**	0.3643	**0.4157**	0.5474	**0.5473**
Multilingual Bert	Resnet50	–	cat	0.5031	0.4650	0.3804	0.3845	0.5277	0.5032
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.5073	0.4810	0.4313	0.3889	0.5557	0.5360
Offensiveness Detection									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.6065	0.6032	0.5772	0.6231	0.6006	0.6448
Multilingual Bert	Vgg16	–	add	0.6343	0.6837	0.6472	0.6389	0.6675	0.7167
Multilingual Bert	Vgg16	Multilingual Bert	add	0.7032	0.7010	0.6740	**0.6621**	0.6822	0.7476
Multilingual Bert	Vgg16	–	cat	0.6093	0.6005	0.5603	0.6613	0.6548	0.7075
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.6555	0.6872	0.6698	0.6615	0.6834	0.7159
Multilingual Bert	Resnet50	–	add	0.6559	0.6912	0.6037	0.6498	0.6859	0.7356
Multilingual Bert	Resnet50	Multilingual Bert	add	0.6862	**0.7071**	0.6482	0.6526	0.7114	**0.7531**
Multilingual Bert	Resnet50	–	cat	0.6405	0.6558	0.6156	0.6331	0.6655	0.7349
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.6475	0.6567	0.6941	0.6526	0.6924	0.7413

Table 6: Recall Results of Different Tasks on MET-Meme

Metaphor Understanding									
				Chinese-English		English		Chinese	
type	text	image	fusion	val	test	val	test	val	test
random	–	–	–	0.5207	0.5207	0.4750	0.4838	0.4610	0.4743
text	Multilingual Bert	–	–	0.7642	0.7480	0.7682	0.7695	0.7645	0.7745
image	–	Vgg16	–	0.6850	0.6611	0.7643	0.7930	0.6252	0.6434
	–	Resnet50	–	0.6728	0.6458	0.7513	0.7604	0.6086	0.6617
Text + Image	Multilingual Bert	Vgg16	add	0.7586	0.7576	0.7930	0.8229	0.7512	0.7496
	Multilingual Bert	Resnet50	add	0.7454	0.7363	0.7995	0.7969	0.7612	0.7761
	Multilingual Bert	Vgg16	cat	0.7691	**0.7607**	0.8034	**0.8333**	0.7678	0.7330
	Multilingual Bert	Resnet50	cat	0.7424	0.7303	0.7930	0.8073	0.7745	**0.7828**
Sentiment Analysis									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.1418	0.1512	0.1163	0.1225	0.1464	0.1555
Multilingual Bert	Vgg16	–	add	0.2987	0.2786	0.2630	0.2474	0.3342	0.3259
Multilingual Bert	Vgg16	Multilingual Bert	add	0.3164	0.3176	0.3021	0.2526	0.3540	0.3350
Multilingual Bert	Vgg16	–	cat	0.3060	0.2713	0.2578	0.2591	0.3375	0.3002
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.3319	0.3207	0.2917	0.2982	0.3681	0.3391
Multilingual Bert	Resnet50	–	add	0.3226	0.3129	0.2682	0.2643	0.3524	0.3424
Multilingual Bert	Resnet50	Multilingual Bert	add	0.3371	0.3150	0.2878	0.2839	0.3912	0.3697
Multilingual Bert	Resnet50	–	cat	0.3127	0.2604	0.2773	0.2708	0.3375	0.3433
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.3403	**0.3254**	0.2969	**0.2995**	0.3970	**0.3714**
Intention Detection									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.1996	0.2013	0.2013	0.1911	0.1911	0.1919
Multilingual Bert	Vgg16	–	add	0.4563	0.4289	0.4115	0.3594	0.5219	0.4839
Multilingual Bert	Vgg16	Multilingual Bert	add	0.4700	0.4497	0.4102	0.4128	0.5335	0.5401
Multilingual Bert	Vgg16	–	cat	0.4482	0.4233	0.3984	0.3555	0.4180	0.3737
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.4675	0.4421	0.4232	0.3984	0.5393	0.5285
Multilingual Bert	Resnet50	–	add	0.4903	0.4634	0.3893	0.4154	0.5500	0.5285
Multilingual Bert	Resnet50	Multilingual Bert	add	0.5041	**0.4807**	0.3900	**0.4275**	0.5575	**0.5608**
Multilingual Bert	Resnet50	–	cat	0.4924	0.4573	0.4023	0.4036	0.5393	0.5261
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.5056	0.4766	0.4375	0.4036	0.5699	0.5476
Offensiveness Detection									
				Chinese-English		English		Chinese	
text	image	metaphor	fusion	val	test	val	test	val	test
–	–	–	–	0.2663	0.2499	0.2375	0.2750	0.2605	0.2465
Multilingual Bert	Vgg16	–	add	0.7546	0.7561	0.7487	0.7096	0.7577	0.7734
Multilingual Bert	Vgg16	Multilingual Bert	add	0.7790	0.7637	0.7526	0.7214	0.7634	0.7816
Multilingual Bert	Vgg16	–	cat	0.7541	0.7739	0.7474	0.7292	0.7577	0.7667
Multilingual Bert	Vgg16	Multilingual Bert	cat	0.7591	**0.7779**	0.7474	0.7448	0.7667	0.7949
Multilingual Bert	Resnet50	–	add	0.7561	0.7612	0.7461	0.7474	0.7601	0.7792
Multilingual Bert	Resnet50	Multilingual Bert	add	0.7734	0.7652	0.7539	**0.7487**	0.7717	0.7998
Multilingual Bert	Resnet50	–	cat	0.7556	0.7644	0.7487	0.7044	0.7618	0.7907
Multilingual Bert	Resnet50	Multilingual Bert	cat	0.7556	0.7734	0.7539	0.7279	0.7709	**0.8015**