SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset

Tan Yue[1,2], Xuzhao Shi[1], Rui Mao[2], Zonghai Hu[1], Erik Cambria[2]*

[1]School of Electronic Engineering, Beijing Key Laboratory of Work Safety Intelligent Monitoring,
Beijing University of Posts and Telecommunications
[2]School of Computer Science and Engineering, Nanyang Technological University
yuetan@bupt.edu.cn, sxzs@bupt.edu.cn, rui.mao@ntu.edu.sg
zhhu@bupt.edu.cn, cambria@ntu.edu.sg

Abstract

Sarcasm poses a challenge in linguistic analysis due to its implicit nature, involving an intended meaning that contradicts the literal expression. The advent of social networks has propelled the utilization of multimodal data to enhance sarcasm detection performance. In prior multimodal sarcasm detection datasets, a single label is assigned to a multimodal instance. Subsequent experiments often highlight the superiority of multimodal models by demonstrating their improvements compared to unimodal models based on these unified labels across multiple modalities. However, our investigation revealed that numerous instances of sarcasm cannot be identified using a single modality. Humans employ the conflict between a statement and factual information as a cue to detect sarcasm, and these cues can stem from different modalities. Then, a unified label for a multimodal instance may be not suitable for the associated text or image. In this work, we introduce SarcNet, a multilingual and multimodal sarcasm detection dataset in English and Chinese, consisting of 3,335 image-text pair samples. We provide annotations for sarcasm in visual, textual, and multimodal data, respectively, resulting in over 10,000 labeled instances. The separated annotation schema for unimodal and multimodal data facilitates a more accurate and reasonable assessment of unimodal and multimodal models.

Keywords: multimodal sarcasm detection, information fusion, benchmark dataset

1. Introduction

Sarcasm can be defined as a form of verbal irony, that involves the use of language to mock, insult, or convey contempt by expressing the opposite of what is meant. Because of its pragmatic effect in conveying opinions and sentiments (Mao et al., 2022, 2023b), sarcasm is extensively utilized in social media and online communication. On the other hand, sarcasm often derives its connotations not solely from textual content, but also from contextual elements and image content (Cambria et al., 2017; Yue et al., 2021; Frenda et al., 2022). Both modalities can provide clues for detecting the conflict between factual information and sarcastic expressions, a crucial aspect for the task of sarcasm detection (Yue et al., 2023).

Such multimodal characteristic renders traditional unimodal analysis methods inadequate, as they overlook the intricate interactions between text and images (Wang et al., 2022). For effective sarcasm detection, it is imperative to integrate multimodal information comprehensively, thereby enhancing the accuracy and robustness of the detection mechanisms. In recent years, a multitude of researchers have devoted significant efforts to develop models to enhance multimodal information fusion (Radford et al., 2021; Yue et al., 2022b; Gandhi et al., 2023), yielding notable advancements.

However, constrained by quality dataset availability, a majority of these studies have relied on a single dataset for experimentation, resulting in potentially inconclusive findings and raising concerns regarding the generalization capabilities of several models (Xu et al., 2020; Pan et al., 2020; Wang et al., 2020; Liang et al., 2022). Within the realm of multimodal sarcasm detection, this paucity of datasets has markedly impeded research progression. On the other hand, we notice that the existing dataset contains only English text, which is insufficient for the realities of social networks that contain multiple languages. Consequently, there has been a growing appeal within the research community for the development of new datasets to propel advancements in the field (Liu et al., 2022).

By revisiting existing multimodal sarcasm detection datasets, we identified some shortcomings in the labeling method. The existing labeling method only defines a unified label for a multimodal instance. The subsequent experiments may misuse the unified label to test both unimodal and multimodal models. In this case, the experiments may assume that a unified multimodal label also works for the associated data from different modalities. However, the incorrect label schema can result in a biased evaluation in multimodal learning. For example, in Table 1, given a multimodal sarcasm detection dataset, empirical studies may misuse multimodal labels assigned to image-text pairs to evaluate their image-only, or text-only baseline models (see the label set in the middle column of Table 1).

---

*Corresponding author

Modalities		Label		New Label
Image		1		0
Text	What a wonderful weather!	1	What a wonderful weather!	0
Multimodal		1		1

Table 1: Comparison of different labeling methods for multimodal sarcasm samples ("1" denotes sarcasm and "0" denotes non-sarcasm).

Nonetheless, it is evident that identifying sarcasm solely based on text or image is inappropriate due to the divergence between the factual information (a picture of a cloudy day) and the statement made (wonderful weather), as these elements exist in separate modalities. Sarcasm becomes evident when considering the combination of the text and image in Table 1, while each individual modality alone does not convey sarcasm. Therefore, the ground truth label set of the given multimodal instance should be $\{0, 0, 1\}$, denoting {image label, text label, multimodal label}, respectively. We also find other cases such as $\{0, 1, 1\}$ or $\{1, 0, 1\}$, which will be investigated in detail in Section 3.3.

Accordingly, we propose SarcNet, a new multilingual and multimodal sarcasm detection dataset, based on a different annotation schema. It contains 3,335 image-text pairs in English (1,270) and Chinese (2,065). We label data by unimodality and multimodalities, separately and obtain 10,005 labels. Each sample is cross-labeled by two independent annotators. The annotators showed substantial or higher levels of agreement in the annotation tasks, as evidenced by Cohen's Kappa values of 0.9032, 0.7129, and 0.9769 for textual, visual, and multimodal labels, respectively.

Well-known unimodal and multimodal models are employed to establish benchmarks on SarcNet. Experiments were conducted with three tasks. For image data, we examined CNN-based (ResNet) and Transformer-based (ViT) models, while text data were evaluated using CNN, RNN, and Transformers. Among these, Transformer-based models such as XLM-RoBERTa, endowed with a larger number of parameters, achieved superior accuracy (76.88%) and F1-score (73.36%). Multimodal models facilitate enhanced integration of image and text information, resulting in improved performance, with DT4MID leveraging XLM-RoBERTa attaining the best results (accuracy of 82.73% and F1-score of 84.40%).

Our main contributions are summarized as follows:

- A new multilingual and multimodal sarcasm detection dataset (SarcNet) is proposed. We collected 3,335 image-text pair samples and delivered 10,005 labels. To the best of our knowledge, SarcNet is the first multilingual and multimodal sarcasm detection dataset.[1]

- We revisit the labeling schema employed for the multimodal sarcasm dataset and identify several shortcomings. In our approach, each sample within the dataset is labeled distinctly for different modalities. Through a meticulous evaluation and analysis of the dataset's labeling schema, we aim to bolster interpretability for the task of multimodal sarcasm detection.

- We perform extensive experiments and establish benchmarks on the dataset using well-known baseline models.

2. Related Works

2.1. Multimodal Sarcasm Detection

Multimodal sarcasm detection combines techniques from natural language processing and computer vision and has attracted increasing attention in recent years. The difficulty of the sarcasm detection task lies in recognizing the interaction between textual and visual clues to identify sarcasm accurately. At the very beginning, sarcasm detection methods predominantly focused on textual data (Riloff et al., 2013), utilizing techniques such as sentiment analysis, contextual embeddings, and linguistic feature extraction (Ghosh and Veale, 2016; Baziotis et al., 2018). Models like BERT (Devlin et al., 2019) have been foundational in achieving significant progress in this domain.

---

[1] https://github.com/yuetanbupt/SarcNet.

However, with the emergence of social media platforms where text often coexists with images and the popularity of multimodal affective computing (Fan et al., 2024), there is a need to explore multimodal approaches for sarcasm detection. A pivotal study by Schifanella et al. (2016) introduced a novel approach combining visual and textual information, showcasing the potential for improved sarcasm detection. Cai et al. (2019) presented a hierarchical fusion model. Subsequent studies have further explored the synergistic effects of text and images in expressing sarcasm.

The development of models capable of processing both text and images, such as CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023), has provided more approaches for multimodal sarcasm detection. These models utilize the complementary nature of information in visual and textual data to enable a more comprehensive understanding of sarcastic expressions. By pre-training on a large amount of multimodal data, these models can get better joint representation and semantic alignment of image-text information for multimodal information fusion prediction.

Additionally, the exploration of attention mechanisms and fusion strategies has been instrumental in refining multimodal sarcasm detection. Attention mechanisms allow models to weigh the importance of different modalities, while fusion strategies enable the effective integration of multimodal features. Pan et al. (2020) introduced a BERT-based model, incorporating inter-modality attention to discern inter-modality incongruity, with an emphasis on both intra and inter-modality incongruities. Similarly, Wang et al. (2020) developed a 2D-Intra-Attention mechanism to extract the relationships between text and images. DT4MID (Tomás et al., 2023) integrated textual and visual transformers within a deep neural network framework to achieve superior multimodal information fusion and representation. Malik et al. (2023) conducted an analysis to ascertain the necessity of image information in comprehending the sarcastic intent embedded within the text. Research in this area has sought to optimize the balance between text and image, ensuring that the subtleties of sarcasm are not lost during fusion.

2.2. Dataset Limitation and Motivation

However, there are very few high-quality datasets in the field of multimodal sarcasm detection. To the best of our knowledge, Most current multimodal sarcasm detection models (Pan et al., 2020; Wang et al., 2020; Liang et al., 2021; Liu et al., 2022) are tested using only one dataset. The dataset was developed by Cai et al. (2019), containing image and text data from Twitter.

The scarcity of datasets has significantly hindered progress in research, a concern echoed by researchers (Liu et al., 2022), who have advocated for the introduction of new datasets to propel advancements in the field. Additionally, the limitation of single-language textual data and inappropriate labeling schema restrict evaluation and interpretability in multimodal sarcasm detection. Accordingly, we present SarcNet, a novel multilingual and multimodal sarcasm detection dataset, comprising image-text pairs in English and Chinese. Each modality within a sample is labeled distinctly. We anticipate that our innovative labeling schema and the quality annotation of the SarcNet dataset will significantly advance sarcasm detection research.

3. SarcNet-Dataset

In this section, we discuss the methodology for data collection and annotation, accompanied by an exhaustive analysis of the dataset.

3.1. Dataset Collection and Annotation

We collected 3,335 samples from WeiBo and Twitter in total. The data were gathered by searching topics such as "#irony", "#sarcasm", and "# 讽刺" (sarcasm in Chinese). Also, we conducted keyword searches targeting terms we deemed closely associated with sarcasm like "卷王" (refers to someone who obsessively focuses on studying or working to the exclusion of all else, often in a competitive and intensive manner). Each sample contains an image and a piece of text. Recognizing the limitations of prevailing labeling approaches, we distinctly labeled the image data, text data, and multimodal data (image-text pairs), resulting in a total of 10,005 labels. To ensure the annotation quality, each sample was annotated by two independent annotators. If both annotators provide a unanimous label in an annotation task, that label becomes the ground truth. In cases where there is no consensus between the two annotators, a third annotator will be brought in to provide annotations. The ground truth is determined by the majority-agreed label.

Annotators were either English or Chinese native speakers with bachelor's degrees, possessing deep understanding of linguistic intricacies, cultural contexts, and societal nuances in the relevant language to effectively detect sarcasm. Thus, we specifically seek annotators below the age of 35, as this demographic is typically well-versed in Internet language expressions. During the annotation process, we showed annotators independent text, image or the combination of text and image in different annotation tasks.

	Annotator2-Y	Annotator2-N
Annotator1-Y	YY	YN
Annotator1-N	NY	NN

Table 2: Labeling consistency statistics for two annotators.

This helps to ensure that each annotation task is not impacted by the former tasks. The label set includes sarcasm and non-sarcasm. The annotators annotated data by the following criteria:

Text Sarcasm: Identify instances where the text expresses a meaning that is opposite to its literal interpretation, often characterized by irony, mockery, or exaggeration.

Image Sarcasm: Recognize images that present a meaning contrary to their literal representation, often associated with humor or satirical elements conveyed through visual cues.

Multimodal Sarcasm: For image-text pairs, detect sarcasm by considering both the textual and visual elements, identifying when the combination implies a meaning opposite to the literal interpretation.

The instructions were written in English and Chinese for the annotators with different language backgrounds, respectively.

3.2. Quality Control

To assess the quality of our data annotation, we employ Cohen's Kappa Statistic, a measure utilized to quantify the level of agreement between two raters or annotators classifying items into mutually exclusive categories. The primary objective of Cohen's Kappa is to determine whether the agreement between annotators on the classification or labeling of a dataset extends beyond mere chance. This metric considers the possibility of random agreement and contrasts it with the actual observed agreement, providing a more nuanced understanding of annotator concordance. The formula for Cohen's kappa is calculated as:

$$\mathcal{K} = \frac{P_o - P_e}{1 - P_e}, \tag{1}$$

where P_o is relative observed agreement among annotators. P_e is the hypothetical probability of chance agreement. Table 2 shows how we calculated the annotation consistency of the two annotators. YY denotes that both annotators' labels for the same sample are positive. YN denotes that for the same sample, Annotator 1's label is positive and Annotator 2's label is negative. NN represents cases where both labels are negative.

Cohen's Kappa	Interpretation
$\mathcal{K} \leq 0$	No agreement
$0 < \mathcal{K} < 0.20$	Slight agreement
$0.20 \leq \mathcal{K} < 0.40$	Fair agreement
$0.40 \leq \mathcal{K} < 0.60$	Moderate agreement
$0.60 \leq \mathcal{K} < 0.80$	Substantial agreement
$0.80 \leq \mathcal{K} < 1$	Near perfect agreement
1	Perfect agreement

Table 3: The interpretation of different values for Cohen's Kappa.

NY signifies instances where Annotator 1 labels the sample as negative and Annotator 2 as positive. The P_o and P_e are calculated as:

$$P_o = \frac{YY + NN}{YY + YN + NY + NN}, \tag{2}$$

$$P_{yes} = \frac{(YY + YN)(YY + NY)}{(YY + YN + NY + NN)^2}, \tag{3}$$

$$P_{no} = \frac{(NN + NY)(YN + NN)}{(YY + YN + NY + NN)^2}, \tag{4}$$

$$P_e = P_{yes} + P_{no}. \tag{5}$$

Cohen's Kappa coefficient is between -1 and 1, and is usually higher than 0, where 0 signifies no agreement, and 1 denotes perfect agreement between the two annotators. Table 3 provides a comprehensive interpretation of the various values associated with Cohen's Kappa, elucidating the degrees of agreement represented by different kappa coefficients. We achieve Cohen's Kappa by 0.9032, 0.7129 and 0.9769, for textual, visual and multimodal labels respectively, indicating that the annotation tasks have achieved substantial agreement or above.

Figure 1: An example of image data with low annotation consistency.

On the other hand, we also observe that Cohen's Kappa of image sarcasm annotations is relatively lower than other annotation tasks due to the ambiguity of visual cues. For instance, as depicted in Figure 1, an image featuring a smiling lady may be interpreted by different annotators as either conveying happiness or ridicule. The absence of additional context may result in varying interpretations by different annotators, thereby complicating the labeling process.

3.3. Dataset Analysis

The dataset encompasses a total of 3,335 multilingual samples, where 2,065 instances are in Chinese, and 1,270 instances are in English (see Table 4). Each sample contains image and text data. We label the image data, text data, and multimodal data (image-text pair), separately, and deliver 10,005 labels in total. As shown in Table 5, we allocate 60%, 20%, 20% of the dataset for training, validation, and testing.

Language	Train	Val.	Test	Total
Chinese	1242	424	399	2065
English	756	247	267	1270
Total	1998	671	666	3335

Table 4: Chinese data and English data statistics.

As we argued in Section 1, in addition to the case of $\{0, 0, 1\}$ ({image label, text label, multimodal label}), we found many other cases, such as $\{0, 1, 1\}, \{1, 0, 1\}$. The examples with $\{0, 1, 1\}$ and $\{1, 0, 1\}$ label distributions are showed in Table 6. In the case with $\{0, 1, 1\}$, when analyzing solely the image data, we may not sufficiently detect sarcasm cues to categorize it as sarcasm. However, with the presence of textual context, identifying sarcasm in textual data becomes more feasible, because it contains both factual information, e.g., "can't even cross my feet" and sarcastic statement, e.g., "Soooooooo much Leg space". Integrating this textual information with the image data can enhance the explicitness of detecting sarcasm. On the other hand, for the case with $\{1, 0, 1\}$ label set, the text data ("lol") contains less information and just means "laugh out loud". However, we can get sarcastic information from the image because the sign (in the upper left corner) states "NO CAMPING, NO FIRES, NO DUMPING", while the fact is that the scene of the picture is cluttered with camping-related waste. Thus, the sarcastic image aids the multimodal sarcasm detection in this case.

As shown in Table 7, for multimodal samples labeled as non-sarcastic (where the multimodal label is 0), most of the associated image and text

Mod.	Class	Train	Val.	Test	Total
Text	Sar.	864	274	308	1446
	Non-Sar.	1134	397	358	1889
	Both	1998	671	666	3335
Image	Sar.	623	224	221	1068
	Non-Sar.	1375	447	445	2267
	Both	1998	671	666	3335
M-mod.	Sar.	1116	372	387	1875
	Non-Sar.	882	299	279	1460
	Both	1998	671	666	3335
Total		5994	2013	1998	10005

Table 5: Label statistics for three different modal data. M-mod. denotes multimodalities.

labels are also 0. However, when considering multimodal samples as sarcastic, only 717 instances, comprising 38% of the total with positive multimodal labels, exhibit both sarcastic text and image labels. This suggests a substantial portion of unimodal evaluation labels would be inaccurate if solely derived from annotations of multimodal data. The inconsistency highlights the need to develop a multimodal sarcasm detection dataset with independent annotation schema for images, text and image-text pairs.

By the statistics in Table 7, a noticeable trend emerges: the majority of sarcastic cues, totaling 716 instances, are identifiable from textual data when only a single modality is marked as sarcastic in image-text pairs. Additionally, there are 127 instances which both images and text are individually non-sarcastic, yet their combination conveys sarcasm. This observation underscores the necessity for a robust sarcasm detection system that can effectively navigate these complexities, e.g., modeling the semantic contrast from different modalities (Mao et al., 2019; Yue et al., 2023).

Finally, there are a few cases with non-sarcastic multimodal labels and sarcastic unimodal labels such as $\{0, 1, 0\}$ or $\{1, 0, 0\}$. When there exists information about sarcasm in the image but the textual content has stated that it is being sarcastic, we do not consider this to be an example of expressing sarcasm. This is because, by definition, there is no conflict between the factual information and the statement from different modalities, e.g., "I asked MemeCreator to generate a sarcastic image" (non-sarcastic text) with a sarcastic image ($\{1, 0, 0\}$). The instances with $\{0, 1, 0\}$ labels are typically due to the very weak connections between text and images. As a result, their combinations cannot be identified as sarcasm.
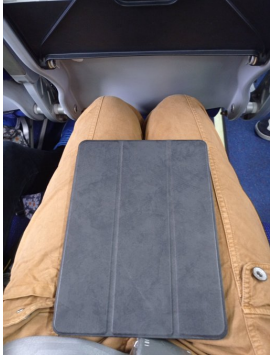
Modalities		Label		Label
Image		0		1
Text	Thank God for Soooooooo much Leg space I can't even cross my feet, let alone my legs.	1	Lol.	0
Multimodal		1		1

Table 6: Examples of $\{0, 1, 1\}$ (image label is 0, text label is 1, multimodal label is 1) and $\{1, 0, 1\}$ (image label is 1, text label is 0, multimodal label is 1).

	$\{0,0,1\}$	$\{1,0,1\}$	$\{0,1,1\}$	$\{1,1,1\}$	$\{0,0,0\}$	$\{1,0,0\}$	$\{0,1,0\}$
Image Label	0	1	0	1	0	1	0
Text Label	0	0	1	1	0	0	1
Multimodal Label	1	1	1	1	0	0	0
Number	127	315	716	717	1411	36	13

Table 7: Different combinations of labels for three modal data.

4. Experiment

We assess our dataset utilizing a variety of well-known unimodal and multimodal models.

4.1. Models

4.1.1. Image-modality methods

ResNet (He et al., 2016), or Residual Network, is a deep convolutional neural network architecture that introduces residual connections to address the vanishing gradient problem and enhance training of deeper networks.

MobileNetV3 (Howard et al., 2019), improved from MobileNetV2, is an efficient convolutional neural network architecture optimized for mobile and edge devices.

ViT (Dosovitskiy et al., 2021)(Vision Transformer) is a model that applies transformer architectures to computer vision tasks by partitioning images into fixed-size patches, then linearly embedding and processing them like a sequence.

4.1.2. Text-modality Methods

LSTM (Hochreiter and Schmidhuber, 1997)(Long Short-Term Memory) is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequences by using memory cells and gating mechanisms.

TextCNN (Kim, 2014) is a convolutional neural network model for text classification, leveraging multiple filter sizes to capture various n-gram features from embedded word sequences.

BERT (Devlin et al., 2019) is a pre-trained transformer-based model designed to understand the context of words in a sentence by considering both left and right surroundings. We also use **MultiL-BERT**, which is a pre-trained model, learning Wikipedia text with 104 languages via a masked language modeling (MLM) objective.

XLM-RoBERTa (Conneau et al., 2020) is an enhanced version of RoBERTa pre-trained on multiple languages, aiming to achieve state-of-the-art performance on cross-lingual NLP tasks.

The LSTM and TextCNN models do not use pre-trained embeddings. Thus, they can learn embed-

Modality	Method	Acc(%)	Pre(%)	Rec(%)	F1(%)
Image	ResNet (He et al., 2016)	67.32	50.16	62.58	55.69
	MobileNetV3 (Howard et al., 2019)	68.37	53.65	56.56	55.07
	ViT (Dosovitskiy et al., 2021)	68.62	53.40	60.28	56.63
Text	LSTM (Hochreiter and Schmidhuber, 1997)	68.47	66.55	63.96	65.23
	TextCNN (Kim, 2014)	70.67	68.08	67.16	67.62
	BERT (Devlin et al., 2019)	73.57	73.74	66.56	69.97
	MultiL-BERT (Devlin et al., 2019)	74.92	76.01	66.88	71.16
	XLM-RoBERTa (Conneau et al., 2020)	76.88	78.52	68.83	73.36
Multimodal	Res-BERT (Pan et al., 2020)	79.48	83.45	80.40	81.89
	KnowleNet (Yue et al., 2023)	80.38	85.06	79.85	82.37
	DT4MID(BERT) (Tomás et al., 2023)	80.53	80.35	84.82	82.52
	DT4MID(XLM-R) (Tomás et al., 2023)	82.73	88.86	80.36	84.40

Table 8: Experimental results (accuracy and F1-score) of unimodal and multimodal models on SarcNet dataset.

dings for both English and Chinese during training. Even though the original BERT is not tailored for multilingual processing, it can also embed and learn tokens from English and Chinese.

4.1.3. Multimodal Methods

Res-Bert (Pan et al., 2020) concentrates on both intra and inter-modality incongruity and is proposed for the sarcasm detection tasks and achieve satisfactory performance.

KnowleNet (Yue et al., 2023) builds cross-modal semantic similarity detection modules and introduces a contrastive learning loss function to optimize the joint representation of multimodal information.

DT4MID (Tomás et al., 2023) integrates textual and visual transformers within a deep neural network framework to achieve superior multimodal information representation.

4.2. Settings

For these unimodal and multimodal models, we use the default parameter settings from their papers. We utilize the pre-trained model as a feature extractor for binary classification tasks. The cross-entropy loss function is employed as the loss function for all models. We train the model with Adam (Kingma and Ba, 2017) optimizer and stop training if the validation loss does not decrease for 5 consecutive epochs. The learning rate is $1e-4$.

4.3. Evaluation

To assess our dataset and establish benchmarks, we utilize **Accuracy**, **Precision**, **Recall**, and **F1-score** as metrics to evaluate.

True Positives (TP): the number of samples for which both the predicted and actual labels are positive.
False Positives (FP): the number of samples with positive predicted labels and negative actual labels.
True Negatives (TN): the number of samples for which both the predicted and actual labels are negative.
False Negatives (FN): the number of samples with negative predicted labels and positive actual labels.

$$Precision = \frac{TP}{TP + FP}, \quad (6)$$

$$Recall = \frac{TP}{TP + FN}, \quad (7)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}, \quad (8)$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \quad (9)$$

5. Results

We conduct experiments using a number of well-known models to evaluate our dataset and analyze the experimental results.

5.1. Main Results

As shown in Table 8, we divide the experiments into three groups and use different labels (image label, text label, and multimodal label) for training and testing.

5.1.1. Text-modality Methods

To assess the text data in our dataset, we employ several well-known models, including CNN-

Modality	Method	Acc(%)	Pre(%)	Rec(%)	F1(%)
Text	BERT (Devlin et al., 2019)	68.92	63.35	61.17	62.24
	MultiL-BERT (Devlin et al., 2019)	70.68	70.33	62.76	66.33
	XLM-RoBERTa (Conneau et al., 2020)	71.93	69.70	66.50	68.06
Multimodal	DT4MID(BERT) (Tomás et al., 2023)	72.43	75.41	67.98	71.50
	DT4MID(XLM-R) (Tomás et al., 2023)	76.19	74.32	81.28	77.65

Table 9: Experimental results of unimodal and multimodal models trained and tested based only on Chinese data.

Modality	Method	Acc(%)	Pre(%)	Rec(%)	F1(%)
Text	BERT (Devlin et al., 2019)	78.03	87.39	69.78	77.60
	MultiL-BERT (Devlin et al., 2019)	80.56	85.86	75.23	80.19
	XLM-RoBERTa (Conneau et al., 2020)	82.65	86.89	76.26	81.23
Multimodal	DT4MID(BERT) (Tomás et al., 2023)	91.76	90.53	93.48	91.98
	DT4MID(XLM-R) (Tomás et al., 2023)	93.96	93.65	96.20	94.91

Table 10: Experimental results of unimodal and multimodal models trained and tested based only on English data.

based models such as TextCNN, RNN-based models such as LSTM, and various transformer-based methods, for assessments. Recent research has predominantly centered on transformer-based approaches such as BERT and its variants, which have demonstrated superior performance. BERT, evaluated solely on text data, achieves an accuracy of 73.57% and an F1-score of 69.97%, underscoring its exceptional efficacy in sarcasm detection. Given the bilingual nature of our dataset, encompassing both Chinese and English, we also incorporate the MultiL-BERT model, pre-trained on the 104 most linguistically diverse languages available on Wikipedia. When compared to BERT, MultiL-BERT exhibits an improvement of 1.35% in accuracy and 1.19% in F1-score, illustrating the effectiveness of multilingual pre-training. The XLM-RoBERTa model, endowed with the highest number of parameters, achieves the optimal accuracy of 76.88% and F1-score of 73.36%.

5.1.2. Image-modality Methods

Additionally, we employ CNN-based models such as ResNet and MobileNetV3, along with the transformer-based model ViT, to evaluate our image data. Among these, ViT achieves the highest accuracy (68.62%) and F1-score (56.63%). However, compared with text-modality methods, the image-modality methods do not perform well, suggesting the difficulty of visual sarcasm detection. This is in line with our human annotations, e.g., Cohen's Kappa of image sarcasm annotations is lower than that of text.

5.1.3. Multimodal Methods

We employ several well-known multimodal models specifically designed for the sarcasm detection task to evaluate our multimodal data. DT4MID achieves the accuracy of 80.53% and F1-score of 82.52%. Given the continuous advancements in pre-trained models for text processing, we observe that the XLM-RoBERTa model achieves the highest results in experiments focused solely on text data. Consequently, we incorporate the XLM-RoBERTa model as the backbone in the DT4MID model. The enhanced performance of XLM-RoBERTa contributes to DT4MID achieving a higher accuracy (82.73%) and F1-score (84.40%).

5.2. Monolingual training results

As shown in Table 4, SarcNet contains multilingual data. To evaluate the monolingual data, we train and test the model using English data and Chinese data, separately. From the experimental results (see Tables 9 and 10), we observe that the overall results for the English data are better than the Chinese data. We consider that this is because sarcastic clues are easier to capture in English language expressions compared to Chinese. The recognition of Chinese sarcastic expressions is notably influenced by Chinese language conventions, rendering it more intricately intertwined with cultural context. Consequently, the recognition of sarcasm in the Chinese language is often perceived as a more challenging task.

5.3. Case study

In our experiments, we found it difficult to predict the cases of $\{0, 0, 1\}$ and $\{1, 0, 0\}$. Fig. 2 is an example of a sarcastic text-image pair with the label of $\{0, 0, 1\}$. The text was designed for an AI to generate an image of the renowned entrepreneur, Jack Ma. However, the AI produced an image depicting a horse and a mountain enveloped by clouds. This is because the literal meaning of Jack Ma ("马云") in Chinese is "horse" ("马") and "cloud" ("云"). This case conveys both the factual content (in text form) and the erroneous image to signify the presence of sarcasm. Nonetheless, it is difficult to capture sarcastic cues between text and image and the examined models likely make incorrect predictions, e.g., $\{0, 0, 0\}$.

Figure 2: Text content: 让 # 文心一言 # 画个马云 (Let #ERNIE Bot# draw the Jack Ma). A difficult case for detection.

6. Conclusion

Sarcasm, as an advanced pragmatic form in natural language, serves as a significant means for humans to convey emotions. However, the recognition of sarcasm presents numerous challenges to machine intelligence, as it requires interpreting information that goes beyond the literal meaning of language (Cambria et al., 2023; Mao et al., 2023a). In this paper, we propose SarcNet, a novel multilingual and multimodal sarcasm detection dataset, comprising 3,335 image-text pair samples and yielding over 10,000 labels. We also revisit the annotation schema employed for multimodal sarcasm datasets, identifying several shortcomings and ensuring that each modality within a sample is labeled distinctly in our dataset.

The distinct image and text labels prove advantageous for more effectively testing unimodal models. The substantial values of Cohen's Kappa demonstrate the substantial agreement of our annotation tasks. Finally, we conduct extensive experiments and establish benchmarks on SarcNet using a range of well-known baseline models, providing a robust foundation for future research in this domain.

8. Ethics Statement

The collected image and text data was anonymized to protect user privacy. We recognized that social groups from different cultures may have different understandings of sarcasm and humor, and that the content in the dataset may be misunderstood or used for inappropriate purposes. Considering the potential impact of the dataset for society, we assessed the culturally sensitive content that may be contained in the dataset and manually filtered sensitive content.

9. Bibliographical References

Christos Baziotis, Nikos Athanasiou, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 3: Tracking ironic Tweets using ensembles of word and character level attentive RNNs. *arXiv preprint arXiv:1804.06659*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of Artificial Intelligence. *IEEE Intelligent Systems*, 38(6):62–69.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

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

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *arXiv preprint arXiv:2210.03501*.

Manuj Malik, David Tomás, and Paolo Rosso. 2023. How challenging is multimodal irony detection? In *International Conference on Applications of Natural Language to Information Systems*, pages 18–32. Springer.

Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. MetaPro: A computational metaphor processing model for text pre-processing. *Information Fusion*, 86-87:30–43.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. 2023a. MetaPro Online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, volume 3, pages 127–135, Toronto, Canada. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3888–3898.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023b. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

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. 2021. Learning transferable visual models from natural language supervision.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 1136–1145.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410.

Heng Wang, Tan Yue, Xiang Ye, Zihang He, Bohan Li, and Yong Li. 2022. Revisit finetuning strategy for few-shot learning to transfer the emdeddings. In *The Eleventh International Conference on Learning Representations*.

Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. 2020. Building a Bridge: A method for image-text sarcasm detection without pretraining on image-text data. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 19–29.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic Tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786.

Tan Yue, Zihang He, Chang Li, Zonghai Hu, and Yong Li. 2022a. Lightweight fine-grained classification for scientific paper. *Journal of Intelligent & Fuzzy Systems*, 43(5):5709–5719.

Tan Yue, Yong Li, and Zonghai Hu. 2021. Dwsa: An intelligent document structural analysis model for information extraction and data mining. *Electronics*, 10(19):2443.

Tan Yue, Yong Li, Xuzhao Shi, Jiedong Qin, Zijiao Fan, and Zonghai Hu. 2022b. Papernet: A dataset and benchmark for fine-grained paper classification. *Applied Sciences*, 12(9):4554.

Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100:101921.