DocMSU: A Comprehensive Benchmark for Document-Level Multimodal Sarcasm Understanding

Hang Du[1], Guoshun Nan[1*], Sicheng Zhang[1], Binzhu Xie[1], Junrui Xu[1], Hehe Fan[2], Qimei Cui[1], Xiaofeng Tao[1], Xudong Jiang[3]

[1]Beijing University of Posts and Telecommunications, China
[2]Zhejiang University
[3]Nanyang Technological University, Singapore
{7597892, nanguo2021, zhangsicheng, xbz_nicous, bbxst0371}@bupt.edu.cn, crane.h.fan@gmail.com, {cuiqimei, taoxf}@bupt.edu.cn, EXDJiang@ntu.edu.sg

Abstract

Multimodal Sarcasm Understanding (MSU) has a wide range of applications in the news field such as public opinion analysis and forgery detection. However, existing MSU benchmarks and approaches usually focus on sentence level MSU. In document level news, sarcasm clues are sparse or small and are often concealed in long text. Moreover, compared to sentence level comments like tweets, which mainly focus on only a few trends or hot topics (e.g., sports events), content in the news is considerably diverse. Models created for sentence level MSU may fail to capture sarcasm clues in document level news. To fill this gap, we present a comprehensive benchmark for Document level Multimodal Sarcasm Understanding (DocMSU). Our dataset contains 102,588 pieces of news with text image pairs, covering 9 diverse topics such as health, business, etc. The proposed large-scale and diverse DocMSU significantly facilitates the research of document level MSU in real world scenarios. To take on the new challenges posed by DocMSU, we introduce a fine grained sarcasm comprehension method to properly align the pixel level image features with word level textual features in documents. Experiments demonstrate the effectiveness of our method, showing that it can serve as a baseline approach to the challenging DocMSU.

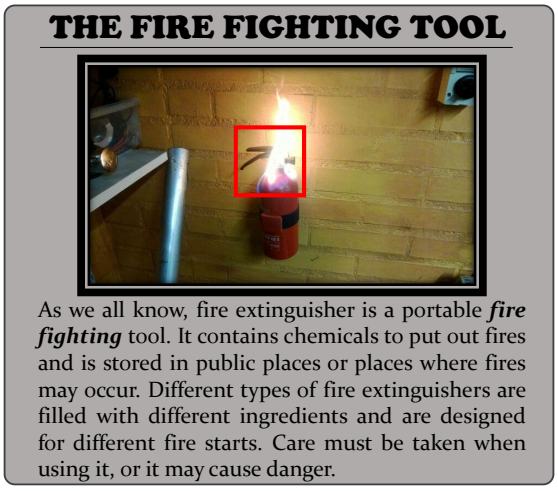

Figure 1: Example of a piece of sarcastic news. It shows that a fire extinguisher, a tool used to extinguish fires, has caught fire. This sarcasm can remind us of the quality of the fire extinguisher.

Introduction

Sarcasm is a form of verbal irony that often uses positive words to convey a negative message, such as frustration, anger, contempt and even ridicule (Wilson 2006). In real-world cases, a piece of sarcastic news often lacks explicit linguistic markers, and thus requires additional cues to reveal the true intentions. The accompanying visual information provides helpful cues to better perceive ironic discrepancies. Multimodal sarcasm (Wang et al. 2022; Shu et al. 2017) is omnipresent in social media posts, forum discussions, and product reviews, and hence the multimodal sarcasm understanding is of great significance for a wide range of applications in the news field such as sentiment analysis (Mao et al. 2021), fake news detection (Ying et al. 2022; Qi et al. 2023), and public opinion analysis.

Figure 1 illustrates a piece of multimedia sarcastic news. To understand this sarcastic news, a model must capture textual cues from multiple sentences, including *fire fighting* and *cause danger*. The accompanying visual cue, *fire on a fire extinguisher* in the figure below, plays an important role in this sarcasm. The figurative and creative nature of such multimodal sarcasm poses a great challenge to the effective perception of the true intention under the guise of overt positive surface involved in a whole document and an image. This requires the design and development of document-level multimodal sarcasm understanding (MSU) methods that specifically take the characteristics of such an ironic expression into consideration.

Prior research has underscored the critical significance of utilizing extensive, high-quality, and challenging benchmarks for the development and evaluation of state-of-the-art deep learning methods across various natural language processing (NLP) tasks (Baltrušaitis, Ahuja, and Morency 2019; Li et al. 2020). In this context, existing sarcasm bench-

*Guoshun Nan is the corresponding author.

marks (Castro et al. 2019; Wang et al. 2022) have demonstrated considerable promise. However, when addressing document-level multimodal sarcasm understanding in the real-world news domain, they exhibit certain limitations, including (1) Limited length of text. In real-world scenarios, a piece of news may include more than 70 words across multiple sentences (Shu et al. 2017), concealing ironic discrepancies beyond sentence boundaries. However, samples in existing multimodal sarcasm datasets (Cai, Cai, and Wan 2019; Castro et al. 2019; Wang et al. 2022) only include about 20 words within a single utterance on average, which greatly simplifies the challenges of MSU in real-world cases. (2) Limited quality of annotations. Existing large satirical datasets (Riloff et al. 2013; Ptáček, Habernal, and Hong 2014; Barbieri, Saggion, and Ronzano 2014) are mostly generated by bootstrapping algorithm or remote supervision with noisy labels. These annotations can be disruptive to systems that harness such data for downstream applications due to the subtle nature of sarcasm. Furthermore, these datasets only contain text modality. (3) Very limited number of samples. As sarcasm lacks explicit linguistic or visual markers, a model requires a large volume of samples to learn the rules or ways that reveal the true underlying intentions. A large-scale dataset benefits the generalization capability of an MSU model that alleviates the over-fitting issue during the training procedure.

The aforementioned limitations in existing datasets highlight the need for a comprehensive, challenging, and higher-quality document-level multimodal sarcasm dataset to enhance irony understanding in the domain of news. Towards that, we developed DocMSU, a comprehensive benchmark that contains high-quality annotations of $102,588$ pieces of news with text-image pairs, covering 9 hot topics such as science, business, and sports. We collect these samples from social websites, including "New York Times" and "UN News", each involving 63 tokens across 5 sentences on average. To alleviate the ambiguity of sarcasm, we manually annotated these documents and images in 3 rounds with 15 workers, ensuring the annotation quality with confidence scores. Each pair of text-image involves a binary label for sarcasm detection, 2.7 textual spans and visual bounding boxes on average for sarcasm localization.

The proposed DocMSU facilitates the research of multimodal sarcasm perception for real-world applications. It also introduces two new challenges: (1) capturing the nuanced sarcastic clues in two modalities, where the clues are concealed within very few words in a document or a tiny area in an image; (2) aligning the visual and linguistic features for irony understanding, where the incongruity nature of sarcasm requires cross-modal interactions. To fill this gap, we propose a novel sarcasm comprehension method that aims to fuse the pixel-level image features with the word-level textual features of a whole document in a fine-grained manner. Experimental results show the effectiveness of our method. The main contributions of our work can be summarised as follows:

- We curate DocMSU, a new benchmark for document-level multimodal sarcasm understanding in the real-world news field. Compared with existing ones, our dataset is more comprehensive and more challenging with much higher quality annotations.
- We come up with a novel document-level MSU method for sarcasm detection and localization, mitigating the issues in sarcastic cues detection across sentences and across modalities under inconsistent context.
- We conduct extensive experiments on our DocMSU. Results show that the created benchmark enables us to develop and evaluate various deep learning methods for the task of MSU closer to the real-world application.

Related Work

Datasets: Existing sarcasm datasets are mainly collected from Twitter and Reddit and can be roughly categorized into text-based ones (Riloff et al. 2013; Ptáček, Habernal, and Hong 2014; Barbieri, Saggion, and Ronzano 2014; Khodak, Saunshi, and Vodrahalli 2018; Oprea and Magdy 2020), and the multimodal ones (Cai, Cai, and Wan 2019; Castro et al. 2019; Wang et al. 2022). The text-based datasets suffer from noisy labels caused by remote supervision. The most related to our work is MSTI (Wang et al. 2022). However, the texts in MSTI only contain 20 tokens on average, which may not well reflect challenges in the news field. Compared to the existing sarcasm datasets, our DocMSU provides more samples, much longer texts and higher quality annotations towards sarcasm understanding in practice of the real-world news field. Detailed comparisons are available in Table 1.

Methods: Early studies of sarcasm understanding were based on statistical patterns (Riloff et al. 2013; Joshi, Sharma, and Bhattacharyya 2015) and deep learning techniques such as word embeddings and LSTM/CNN (Joshi, Sharma, and Bhattacharyya 2015; Zhang, Zhang, and Fu 2016). Recent MSTI leverages pre-trained BERT and ResNet to extract the cross-modal features(Wang et al. 2022). Some powerful methods such as CLIP (Radford et al. 2021) and VILT(Hu et al. 2019) rely on contrastive learning and Transformer to learn multimodal representations. Different from the above methods, our model aims to comprehend the fine-grained nuanced sarcastic clues in two modalities, where the clues reside within very few words in a document or a very tiny area in an image.

The DocMSU Dataset

We present DocMSU, a new benchmark that contains high-quality annotations of $102,588$ pieces of news with text-image pairs in 9 hot topics.

Data Collection

We crawl data from some famous news websites such as "New York Times", "UN News", "The Onion" and "NewsThumb", etc. To avoid regulation issues, we discard news that includes sensitive topics such as pornography and violence. Finally, we collect more than $70,000$ pieces of news that consist of titles, abstracts, images, and news bodies, where each sample is generated by combing a news title, the abstract, and the image. Each sample involves 63 tokens across 5 sentences on average. We categorize these data into

Datasets	Volume	Level	Source	Input	Labeled Obj.	Annotator
Riloff (Riloff et al. 2013)	175,000	Sentence	Tweets	Text	-	Manual
iSarcasm (Oprea and Magdy 2020)	4,484	Sentence	Social media	Text	Text	Manual
Cai (Cai, Cai, and Wan 2019)	24,635	Sentence	Tweets	Text, Image	Text	Auto
MSTI (Wang et al. 2022)	5,015	Sentence	Tweets	Text, Image	Text, Image	Manual
MUStARD (Castro et al. 2019)	690	Sentence	TV shows	Text, Audio, Video	-	Manual
DocMSU (ours)	**102,588**	**Document**	**News**	Text, Image	**Text, Image**	Semi-Manual

Table 1: The comparisons between our DocMSU dataset and previous ones.

9 groups such as "science", "health", and "business", and each group involves 10 visual object types such as "building", "animal" and "art". We use an open-source tool doccano (Nakayama et al. 2018) for textual and visual annotations and 15 volunteers participated in the work. We mix up different categories of samples before the annotation, allowing each annotator to randomly access news.

Annotation Process

During the annotation procedure, we give a binary tag for each document-image pair to indicate whether it is an ironic message. For a piece of sarcasm news, we further mark the sarcastic clues, including the textual span in the document and the bounding box in the image. However, we face two challenges in such an annotation procedure.

- **Lacking explicit linguistic and visual markers in a sample.** An annotator may not be able to accurately understand the sarcasm in some news titles, images and the corresponding abstracts, as they may require some proper background knowledge for the annotator to understand the sarcasm.
- **Annotation variances caused by the subjective nature of perceiving sarcasm.** As irony is always conveyed in a subtle way both in a document or an image, the perception of sarcastic clues varies from different annotators.

For the first issue, we ask the annotator to refer to the news body to better understand the context. By doing so, the annotator is able to give a more accurate binary label, as well as sarcastic clues including the textual spans in the document and bounding box in the image. Regarding the second issue, we have 3 annotators for each sarcastic sample with a scoring mechanism. We use Intersection-over-Union (IoU) to quantify the similarity between two annotations. A similarity score between two annotations is defined as the sum of textual IoU (TIoU) and visual IoU (Yu et al. 2016). TIoU is defined as follows: S refers to the text labeled by the annotator, r is the index of annotator, i and j indicate the positions of the beginning and the end of the sarcastic span respectively.

$$TIOU = \frac{min(S_{r-1}[j], S_r[j]) - max(S_{r-1}[i], S_r[i])}{max(S_{r-1}[j], S_r[j]) - min(S_{r-1}[i], S_r[i])} \quad (1)$$

For each annotation, we obtain two similarity scores with the other two annotations. The sum of them is defined as the confidence score of this annotation. The annotation with the highest confidence score is selected in our DocMSU. Due to the subtle nature of sarcasm, there are some samples whose ironic clues can be hardly distinguished. For these samples, we observe that all the three confidence scores are much smaller than those of other samples, and these samples take up about 5 percent of the data. Hence, we ask the annotator who achieves the overall highest confidence score among the

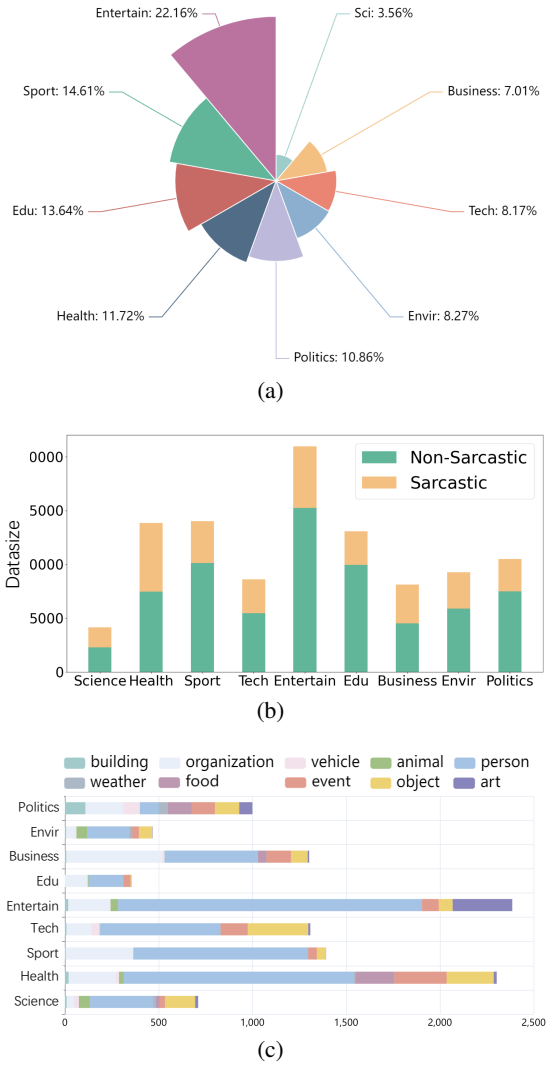

Figure 2: Statistics of our DocMSU: (a) percentage of each topic in the overall dataset. (b) distribution of sarcastic samples and non-sarcasm ones in each news topic. (c) distribution of visual object type in each topic.

15 volunteers in the whole annotation procedure to further label these "challenging" instances. We also use GPT-3.5 to augment the text data and discard instances that may include sensitive information.

Dataset Analysis

Figure 2 details the statistics of our DocMSU. Figure 2(a) shows the percentage of the 9 topics, "Science", "Health", "Sport", "Technology", "Entertainment", "Education", "Business", "Environment", and "Politics", where the "Environment" topic is most popular and takes the largest portion 22.16%. Figure 2(b) illustrates the distribution of sarcastic samples and non-sarcasm ones in each topic. Totally, our benchmark contains 34,130 sarcastic samples and 68,458 non-sarcastic ones. Figure 2(c) shows the distribution of visual object type in each topic, where multiple types of visual objects enrich the feature for sarcasm understanding. We have 10 object types in our DocMSU. A sample contains 2.7 labeling targets on average, which are sarcastic clues, including textual spans in a document and bounding boxes in an image[1].

Figure 3: Two samples selected from our benchmark.

Proposed Method

Motivation

Figure 3 shows two examples selected from our DocMSU. The first example highlights the irony in National Geographic's practices, as they extensively employ plastic packaging despite advocating against the overuse of plastic. The subsequent example exposes the contradictions within Netflix's service delivery. Despite boasting about their discs being "unbreakable" and implying exceptional durability, customers received damaged discs, contradicting the advertised durability.

A model may face two new challenges for sarcasm understanding the above two examples. (1) Capturing the nuanced sarcastic clues that are concealed within very few words (e.g., "*a broken disc*") in a document or in a very tiny area (e.g., the tag "*unbreakable*") of the image. (2) Aligning the visual and text features for the accurate irony understanding (e.g., "*unbreakable*" and the broken disc). Existing approaches such as recent MSTI (Wang et al. 2022), CLIP, and VILT have limitations in tackling these two challenges as they focus more on learning the overall information of the whole text and image representations. This motivates us to develop a new method to capture the fine-grained linguistic and visual sarcastic clues and align the two different types of clues for a better MSU.

Overview

Figure 4 illustrates the architecture of our model, which consists of three components, including a document encoder, an image encoder, and a fusion module. To capture the underlying subtle clues concealed within very few words in a document and a very tiny area in an image, our model generates two matrices for pixels-level image representations and token-level document representations. For the cross-modal interactions, we fuse the representations in two matrices with a sliding window for multimodal alignment. We explore the specifics of this design in the following sections.

Document Encoder

We denote a document as $s = \{w_i\}_{i=1}^{n}$, where w_i indicates the i-th token and n is the total number. We use BERT (Devlin et al. 2019) to output contextualized token-level representations $\mu \in R^{n\times d}$, where

$$\mu = [\nu_1, \nu_2, ..., \nu_n] = \mathrm{BERT}(s) \tag{2}$$

We use a fully connected layer f_c to transform word representations. Then, we convert the document representation into a square shape $\varpi \in R^{L\times L\times d}$,

$$\varpi(i, j, :) = \big(f_c(\theta)\big)_{L\times(i-1)+j} \tag{3}$$

where $1 \le i, j \le L$. We add paddings when $n < L \times L$. This square document representation is used for the fine-grained alignment of the pixel-level visual representations.

Image Encoder

To keep the high spatial resolution of the feature maps and retain the information of the image details, we only use the early three convolution layers of ResNet (He et al. 2016). In this way, we can keep the original resolution of images. Then, we use a projection layer f_p to generate the visual representations for each pixel. Third, the image feature map is spatially divided into m sliding windows, with $L \times L$ pixels for each window. In this way, the representation of the entire image is as follows,

$$\omega = [\omega_1, \omega_2, ..., \omega_m] = f_p\big((\mathrm{ResNet}(\sigma)\big), \tag{4}$$

where σ is the input image and $\omega_k \in R^{L\times L\times d}$ denotes the k-th window.

[1]We provide more details in Appendix: DocMSU Annotation Pipeline, including annotation user interface, data samples, etc. and appendices are available in the preprint version.

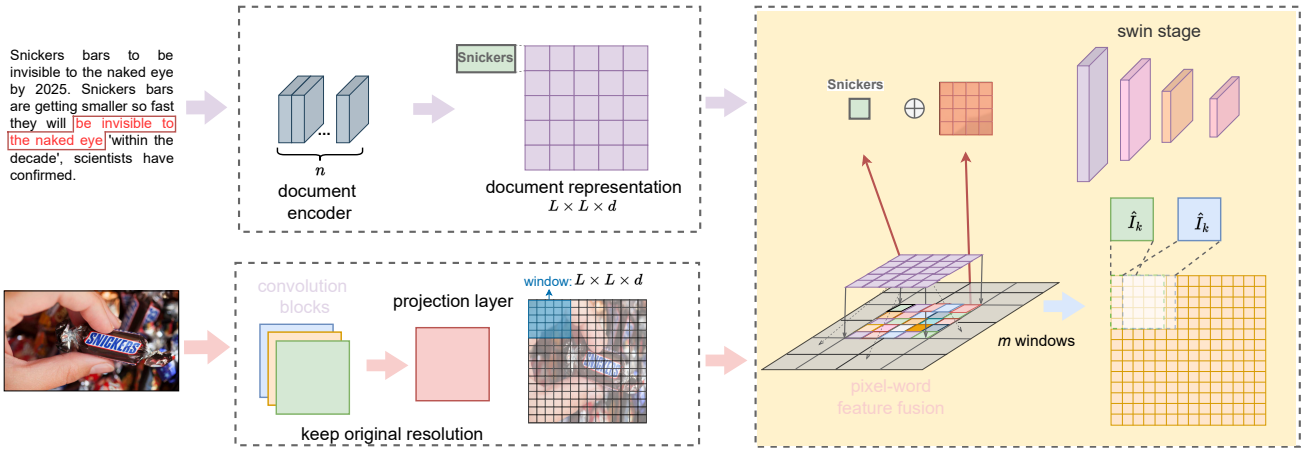

Figure 4: Overview of the proposed model

Multimodal Sarcasm Fusion

During the process of multimodal fusion, we add the document representation ϖ to each window ω_k. The result of addition is denoted as $\hat{\omega}_k$,

$$\hat{\omega}_k = \varpi + \omega_k \quad k = 0, 1, ..., m \tag{5}$$

We apply four stages in Swin-Transformer (Liu et al. 2021) to deeply fuse the two modalities. Specifically, each stage contains one patch merging layer and several blocks containing mechanisms of shifted window attention, which calculates the attention between each element in each shifted window with little computational complexity. By doing so, interactions are built between each word of the document and each image pixel without adding additional calculations. The output of this method delivers multimodal fine-grained features that could be applied to sarcasm understanding.

Experiments

Evaluation Tasks

To evaluate our model, we perform two MSU tasks, i.e., sarcasm detection and sarcasm localization.

Sarcasm detection: Sarcasm detection aims to identify whether visual or verbal irony exists in the sample. This task can be formulated as a binary classification problem.

Sarcasm localization: Sarcasm localization aims to find out the sarcastic clues or objects in a document with textual spans, as well as in the paired image with bounding boxes.

Implementation Details and Settings

We employ the pre-trained uncased BERT-base (Devlin et al. 2019) as the text encoder. For sarcasm localization, we use a linear layer to predict whether a word token is sarcastic in the text and employ YoloX (Ge et al. 2021) as the head network to output the bounding box of the sarcastic object or region. For sarcasm detection and textual sarcastic localization, we use the binary cross entropy loss function. For visual sarcastic localization, we employ the CIoU loss function (Zheng et al. 2022). We train our model with a single NVIDIA RTX 3090 GPU. The learning rate is set to 0.001 and 0.01 for sarcasm detection and localization, respectively. We employ AdamW (Kingma and Ba 2014) as the optimizer. The dataset is randomly split into 70%, 20%, and 10% for training, validation, and testing. The previous Swin-Transformer has three settings including $Tiny$, $Small$, and $Base$ (Liu et al. 2021). For the baseline, we configure Swin-Transformer with the $Tiny$ setting for sarcasm localization, and $Base$ for the detection task, as such two settings perform best among all three settings in corresponding tasks. More details are available in Appendix.

Evaluation Matrices

For sarcasm detection and sarcasm localization in images, we follow (Wang et al. 2022) and (Lin et al. 2014) to use average precision (AP) and F1 scores for evaluation, respectively, including AP_{50}, AP_{60}, $F1_{50}$ and $F1_{60}$. For textual sarcasm localization, Exact Match (EM) (Joshi et al. 2018) is usually employed to measure the prediction accuracy, which is defined as the number of correct predictions that strictly (100%) match the boundaries of annotations divided by the total number of predicted samples. However, as shown in Figure 5, the original EM is too strict to reflect the prediction accuracy. Therefore, we introduce three new evaluation matrices as follows,

- **EM_{50}, EM_{70}:** We use EM_{50} and EM_{70} to relax the standard EM. They are defined as the number of predictions that match more than 50% and 70% annotations divided by the total number of predicted samples. The original EM can be seen as EM_{100}.
- **BitError:** BitError is the ratio of those wrongly classified tokens to the total number of tokens in the sample.

Sarcasm Detection Results

In this paper, we compare our method with BERT-base (text-only) (Devlin et al. 2019), Swin Transformer (image-only) (Liu et al. 2021), CLIP (Radford et al. 2021), Vision-and-Language Transformer (ViLT) (Kim, Son, and Kim 2021) and CMGCN (Liu et al. 2021), which detects sarcasm by the object types. For CLIP and ViLT, we first concatenate

	Sarcasm Localization				
	Model	$AP_{50}\uparrow$	$F1_{50}\uparrow$	$AP_{60}\uparrow$	$F1_{60}\uparrow$
Image	Swin-Transformer (Liu et al. 2021)†	21.78($\pm$0.09)	21.70($\pm$0.10)	6.13($\pm$0.21)	6.10($\pm$0.19)
	MSTI (Wang et al. 2022)	10.21($\pm$0.95)	10.17($\pm$0.94)	6.31($\pm$1.21)	6.29($\pm$1.17)
	CLIP (Radford et al. 2021)	17.76($\pm$0.55)	17.65($\pm$0.53)	6.23($\pm$0.26)	6.23($\pm$0.27)
	ViLT (Kim, Son, and Kim 2021)	30.73($\pm$1.25)	30.68($\pm$1.23)	7.64($\pm$0.99)	7.58($\pm$1.47)
	Ours$_{ST}$	34.17($\pm$2.01)	33.99($\pm$2.12)	10.17($\pm$1.21)	10.05($\pm$1.01)
	Ours$_{SS}$	32.95($\pm$1.79)	32.90($\pm$1.81)	11.68($\pm$1.51)	11.51($\pm$2.22)
	Ours$_{SB}$	**35.29**($\pm$2.92)	**35.24**($\pm$2.95)	**13.74**($\pm$2.71)	**13.67**($\pm$1.67)
	Model	$EM_{50}\uparrow$	$EM_{70}\uparrow$	EM$\uparrow$	BitError$\downarrow$
Text	BERT-base (Devlin et al. 2019)†	45.86($\pm$0.15)	36.86($\pm$0.26)	35.77($\pm$0.42)	18.24($\pm$0.11)
	MSTI (Wang et al. 2022)	44.99($\pm$0.71)	35.45($\pm$0.73)	34.10($\pm$0.82)	**14.70**($\pm$0.20)
	CLIP (Radford et al. 2021)	39.34($\pm$0.83)	34.14($\pm$0.85)	33.62($\pm$0.99)	20.47($\pm$0.31)
	ViLT (Kim, Son, and Kim 2021)	44.09($\pm$0.81)	36.96($\pm$0.77)	36.13($\pm$0.75)	24.64($\pm$0.93)
	Ours$_{ST}$	49.74($\pm$0.87)	40.17($\pm$1.06)	38.50($\pm$1.04)	17.12($\pm$0.30)
	Ours$_{SS}$	50.68($\pm$0.92)	41.17($\pm$1.59)	**39.91**($\pm$0.97)	16.83($\pm$0.71)
	Ours$_{SB}$	**52.19**($\pm$1.59)	**43.88**($\pm$0.51)	39.66($\pm$0.29)	17.67($\pm$1.16)

Table 2: Comparisons of our method with pre-trained single-modal (denoted as †) and multimodal baselines for sarcasm localization. Here we use the subscripts ST, SS and SB to represent the different settings of Swin-Transformer in our method, including $Tiny$, $Small$ and $Base$ (Liu et al. 2021).

the global image and text features and then perform binary classification. Our model employs three settings of Swin-Transformer to extract image representations respectively.

As shown in Table 3, our method with Swin-Transformer of $Base$ achieves the best accuracy, demonstrating its superiority in sarcasm detection. Because the single-modal

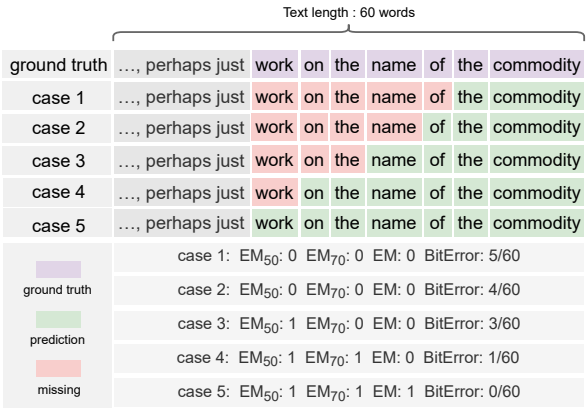

Figure 5: Illustration of the proposed EM_{50}, EM_{70}, and BitError evaluation metrics for textual sarcastic localization, under the assumption that there are five prediction cases. As case 5 makes the completely correct prediction, its original EM (or EM_{100}) can be considered as 1. Although cases 1 to 4 yield partially correct predictions, their EM scores are 0 due to the excessively strict rule of EM. Therefore, the original EM cannot properly measure the accuracy. Conversely, our EM_{50} and EM_{70} can better reflect the accuracy from different levels. We additionally introduce a BitError matrix to evaluate predictions based on errors. The proposed evaluation metrics EM_{50}, EM_{70} and BitError comprehensively reflect the prediction accuracy of MSU.

Sarcasm Detection			
Model	**Acc**	**Pre**	**F1-score**
BERT-base (Devlin et al. 2019)†	87.12	77.61	86.51
Swin-Transformer (Liu et al. 2021)†	74.83	67.57	61.51
CMGCN (Liang et al. 2022)	88.12	78.11	75.23
CLIP (Radford et al. 2021)	96.19	78.99	77.62
ViLT (Kim, Son, and Kim 2021)	93.15	69.03	41.44
Ours$_{ST}$	96.40	76.71	80.16
Ours$_{SS}$	96.82	78.10	82.75
Ours$_{SB}$	**97.83**	**81.20**	**87.25**

Table 3: BERT and Swin-Transformer are based on the single modality (denoted as †). Here the subscripts ST, SS, and SB refer to $Tiny$, $Small$, and $Base$ settings of Swin-Transformer, respectively.

BERT-base and Swin-Transformer do not comprehensively exploit the image and text information, they only achieve suboptimal results. Moreover, the proposed method also outperforms the multimodal CLIP and ViLT models. This is because our method is based on more fine-grained visual signals, and the sliding-window-based Transformer mechanism can better capture the sarcasm clues. More detailed comparisons are available in Appendix: Analysis of Swin-Transformer under different settings.

Sarcasm Localization Results

For the sarcasm localization, we additionally include the sentence-level Multimodal Sarcasm Target Identification (MSTI)(Wang et al. 2022) method, which aims at finding sarcasm clues in tweets. The experimental results are shown in Table2. Our method outperforms those existing methods in both visual and textual sarcasm localization in terms of AP-, F1- and EM-based metrics. For example, in textual sarcasm localization, our method (SB) surpasses CLIP, ViLT,

Sarcasm Detection			
Modality	Accuracy↑	Precision↑	F1-score↑
Image	74.83	67.57	61.51
Text	87.12	77.61	86.51
Image + Text	**97.83**	**81.20**	**87.25**

Table 4: Impact of modalities.

and MSTI by 12.85%, 7.20%, and 8.10%, respectively, in terms of EM_{50}. This shows the effectiveness of the proposed method in localizing nuanced clues in images or long text, and also implies the meaningfulness of collecting such a document-level benchmark. For textual sarcasm localization, we also observe that our method (SS) is the second best and performs slightly lower than MSTI by 2.13 points in terms of BitError. Nevertheless, our method (SS) significantly outperforms MSTI in terms of EM-based metrics, e.g., 5.81 and 7.20 points respectively in terms of EM and EM_{50}. The results suggest that our method can achieve a good balance between coverage and precision during the localization. We will further investigate such an interesting finding in the future. We also provide a case study to visually demonstrate how our model performs multimodal sarcasm localization. Due to the space limitation, we give such an illustration in Appendix: Case Study.

Attention Visualization

Figure 6 depicts the attention map generated by our method. Our model focuses more on the fire extinguisher after four alignment stages, as discussed in section of the multimodal sarcasm fusion. This demonstrates the superiority of our approach in capturing fine-grained textual and image clues in document-level multimodal news.

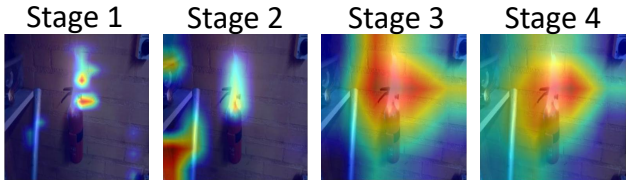

Figure 6: Attention visualization of our method.

Ablation Study

Impact of Modalities on MSU. This section investigates the influence of visual and textual modalities on MSU. As shown in Table 4 and Table 5, combining the two modalities significantly improves the detection and localization accuracy. Such comparisons confirm our hypothesis that multimodal cues can benefit sarcasm understanding at the very beginning of section of the introduction.

Impact of Image-Text Fusion Method. To evaluate the effectiveness of the fusion method, we compare ours with a baseline where the encoded image pixels and text tokens are directly concatenated for sarcasm detection and localization. As shown in Table 6 and Table 7, the proposed fusion method can better capture and align the visual and textual sarcasm clues and achieves better accuracy than the sample concatenation fusion method. These findings show the superiority of our method for the challenging MSU task.

Sarcasm Localization				
Modality	AP_{50}↑	$F1_{50}$↑	AP_{60}↑	$F1_{60}$↑
Image	21.78	21.70	6.13	6.10
Image + Text	**35.29**	**35.24**	**13.18**	**13.12**
Modality	EM_{50}↑	EM_{70}↑	EM↑	BitError↓
Text	45.86	36.86	35.77	18.24
Text + Image	**49.74**	**40.17**	**38.50**	**17.12**

Table 5: Impact of modalities.

Sarcasm Detection			
Fusion Method	Accuracy↑	Precision↑	F1-score↑
Concatenation	90.27	79.69	86.62
The Proposed	**97.83**	**81.20**	**87.25**

Table 6: Effectiveness of our image-text fusion method.

Sarcasm Localization				
Fusion Method	AP_{50}↑	$F1_{50}$↑	AP_{60}↑	$F1_{60}$↑
Concatenation	21.51	21.40	5.31	5.30
The Proposed	**41.04**	**40.87**	**27.08**	**27.01**
Fusion Method	EM_{50}↑	EM_{70}↑	EM↑	BitError↓
Concatenation	43.16	33.33	32.37	18.36
The Proposed	**52.19**	**42.33**	**40.59**	**17.19**

Table 7: Effectiveness of our image-text fusion method.

DocMSU with Large Language Models

We conducted experiments on large language models (LLMs), including GPT-4 (OpenAI 2023), VideoChat (Li et al. 2023b), Otter (Li et al. 2023a), and mPLUG-Owl (Ye et al. 2023). For the instances with obvious satirical clues, LLMs can yield satisfied performance. While for the challenging ones, LLMs still struggle to accurately comprehend sarcasm. Detailed results are presented in Appedix: Tests on LLMs. Particularly, we observe that LLMs encounter difficulty in accurately identifying the satirical object and its underlying cause when the text does not obviously indicate satire. Furthermore, it shows that LLMs excel in providing insightful explanations when the news involves intricate cultural knowledge and social context.

Conclusion

This paper presents DocMSU, a new benchmark for the challenging document-level multimodal sarcasm understanding in the news field. Compared with the existing ones, our DocMSU is more comprehensive, challenging, and involves higher-quality annotations. We believe our DocMSU will encourage the exploration and development of various downstream tasks for document-level multimodal sarcasm perception closer to real-world applications. Future work could focus on MSU across various cultures, as well as the expressive differences between males and females.

Ethical Statement

We have the copyright of contents collected from three websites, including TheOnion, UNNews, and NewsThump, as these sites automatically grant copyright for users who follow their online rules. We carefully study these rules and strictly conform to the requirements during data collection and annotation. These online copyright requirements are available on the above websites. To further fortify ethical compliance, we will take the following steps: 1). Implementing rigorous data anonymization techniques to safeguard personal information. 2). Ensuring transparency about the data sources and collection methods in our revised manuscript. 3). Committing to ongoing scrutiny and readiness to remove or alter data that may be deemed ethically inappropriate or has been collected from sources that do not provide the necessary authorization. 4). Developing an online agreement to require every user of the dataset strictly conform to the rules of the websites from which we collected the data.

Acknowledgments

This work was partially supported by the joint funds for Regional Innovation and Development of the National Natural Science Foundation of China (No. U21A20449), the Beijing Natural Science Foundation under Grant M21037, and the Fundamental Research Funds for the Central Universities under Grant 2242022k60006.

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443.

Barbieri, F.; Saggion, H.; and Ronzano, F. 2014. Modelling Sarcasm in Twitter, a Novel Approach. 50–58.

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. 2506–2515.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4619–4629. Florence, Italy: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 4171–4186.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding YOLO Series in 2021. *arXiv e-prints*, arXiv:2107.08430.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity Mappings in Deep Residual Networks. *ArXiv preprint*, abs/1603.05027.

Hu, H.; Zhang, Z.; Xie, Z.; and Lin, S. 2019. Local Relation Networks for Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3464–3473.

Joshi, A.; Goel, P.; Bhattacharyya, P.; and Carman, M. 2018. Sarcasm Target Identification: Dataset and An Introductory Approach.

Joshi, A.; Sharma, V.; and Bhattacharyya, P. 2015. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 757–762. Beijing, China: Association for Computational Linguistics.

Khodak, M.; Saunshi, N.; and Vodrahalli, K. 2018. A Large Self-Annotated Corpus for Sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5583–5594. PMLR.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; and Liu, Z. 2023a. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 121–137. Cham: Springer International Publishing. ISBN 978-3-030-58577-8.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 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)*, 1767–1777. Dublin, Ireland: Association for Computational Linguistics.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mao, Y.; Shen, Y.; Yu, C.; and Cai, L. 2021. A Joint Training Dual-MRC Framework for Aspect Based Sentiment Analy-

sis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15): 13543–13551.

Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; and Liang, X. 2018. doccano: Text Annotation Tool for Human. Software available from https://github.com/doccano/doccano.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Oprea, S.; and Magdy, W. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1279–1289. Online: Association for Computational Linguistics.

Ptáček, T.; Habernal, I.; and Hong, J. 2014. Sarcasm Detection on Czech and English Twitter. 213–223.

Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14444–14452.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. 704–714.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *Sigkdd Explorations*.

Wang, J.; Sun, L.; Liu, Y.; Shao, M.; and Zheng, Z. 2022. Multimodal Sarcasm Target Identification in Tweets. 8164–8175.

Wilson, D. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10): 1722–1743. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Jiang, C.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2022. Bootstrapping Multi-View Representations for Fake News Detection. In *AAAI Conference on Artificial Intelligence*.

Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; and Huang, T. S. 2016. UnitBox: An Advanced Object Detection Network. *Proceedings of the 24th ACM international conference on Multimedia*.

Zhang, M.; Zhang, Y.; and Fu, G. 2016. Tweet Sarcasm Detection Using Deep Neural Network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2449–2460. Osaka, Japan: The COLING 2016 Organizing Committee.

Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; and Zuo, W. 2022. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on Cybernetics*, 52(8): 8574–8586.