Multi-source Semantic Graph-based Multimodal Sarcasm Explanation Generation

Liqiang Jing[1], **Xuemeng Song**[1]*, **Kun Ouyang**[1], **Mengzhao Jia**[1], **Liqiang Nie**[2]
[1]Shandong University
[2]Harbin Institute of Technology (Shenzhen)
{jingliqiang6, sxmustc, kunouyang10, jiamengzhao98, nieliqiang}@gmail.com

Abstract

Multimodal Sarcasm Explanation (MuSE) is a new yet challenging task, which aims to generate a natural language sentence for a multimodal social post (an image as well as its caption) to explain why it contains sarcasm. Although the existing pioneer study has achieved great success with the BART backbone, it overlooks the gap between the visual feature space and the decoder semantic space, the object-level metadata of the image, as well as the potential external knowledge. To solve these limitations, in this work, we propose a novel mulTi-source sEmantic grAph-based Multimodal sarcasm explanation scheme, named TEAM. In particular, TEAM extracts the object-level semantic meta-data instead of the traditional global visual features from the input image. Meanwhile, TEAM resorts to ConceptNet to obtain the external related knowledge concepts for the input text and the extracted object meta-data. Thereafter, TEAM introduces a multi-source semantic graph that comprehensively characterize the multi-source (*i.e.,* caption, object meta-data, external knowledge) semantic relations to facilitate the sarcasm reasoning. Extensive experiments on a public released dataset MORE verify the superiority of our model over cutting-edge methods.

1 Introduction

Sarcasm is a common linguistic phenomenon, especially in posts on online social media platforms, that expresses people's emotions or opinions in a contrary manner. Since it benefits various real-world applications, such as customer feedback analysis and public opinion analysis, the sarcasm detection task has gained increasing research attention (Joshi et al., 2015; Abercrombie and Hovy, 2016). Despite related great studies of the task, they can only identify the sarcastic post but could not give the concrete explanation for why it is sarcastic, making their detection results less convincing.

Input Image	
	**Input Caption**: shame my city is so ugly.
	Target Explanation: this city is beautiful.
	External Knowledge: ugly->beautiful; luminous->bright

Figure 1: An example of the sarcasm explanation from MORE (Desai et al., 2022). The key objects in the image are marked and the external knowledge is provided.

Noticing this issue, recent studies have shifted to the task of sarcasm explanation, which aims to generate a natural language sentence to explain the intended irony in a sarcastic post. For example, Peled and Reichart utilized the Recurrent Neural Network (RNN) (Ghosh et al., 2017)-based encoder-decoder architecture to tackle the sarcasm interpretation task. Although previous studies have attained impressive results, they focus on investigating the sarcasm explanation purely based on the textual input. Nevertheless, with the advances of multimedia devices, people tend to express their emotions or opinions through multimodal social posts. Moreover, the visual content usually also conveys important clues for explaining the sarcasm, as shown in Figure 1. Motivated by this, Desai et al. proposed the task of multimodal sarcasm explanation, which aims to generate the explanation for a multimodal input (*i.e.,* an image plus its corresponding caption). The authors gave a solution that first fuses the multimodal features with a cross-modal attention module, and then generates the explanation with the decoder of BART, a popular generative pretrained language model. Although this pioneer study has achieved promising performance, it still suffers from three key limitations.

- **L1: Overlook the gap between the visual feature space and the decoder semantic space.** The existing method directly adopts the visual feature of the input image with the

*Xuemeng Song is the corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 11349–11361
July 9-14, 2023

context of BART decoder. In fact, the visual features may not match the semantic space of the BART well since it is pretrained only on the textual corpus, and these existing methods could not maximize the generation capacity of BART.

- **L2: Overlook the object-level metadata of the image.** The existing work only extracts the global feature of the image, ignoring that only the key objects in the image relevant to the input caption contribute to sarcasm explanation (*e.g.,* "luminous building" and "red light" in Figure 1). Moreover, the object's metadata, *e.g.,* the class and attribute, which conveys important clues for the semantic understanding of the visual modality, merits our attention.

- **L3: Overlook the potential external knowledge.** The pioneer study fails to utilize the related knowledge contained in the external public knowledge base. As shown in Figure 1, the related knowledge concepts obtained from ConceptNet (Ghosal et al., 2020) can strengthen the context learning (*e.g.,* bright) and promote the explanation generation (*e.g.,* beautiful).

To tackle these limitations, we propose a novel mulTi-source sEmantic grAph-based Multimodal sarcasm explanation generation scheme, TEAM for short, which explores three semantic sources: the input caption, object meta-data derived from the input image, as well as the external knowledge. Specifically, TEAM includes four components: vision-based object-level semantic extraction, external related knowledge acquisition, multi-source semantic graph-based sarcasm reasoning, and sarcasm explanation generation. As shown in Figure 2, in the first module, we focus on extracting the semantic meta-data of the key objects in the input image instead of the conventional global visual features, to adapt the decoding space of BART and facilitate the fine-grained sarcasm reasoning. In the second module, we target at acquiring the external related knowledge concepts for the input caption and the extracted object meta-data, where a large-scale knowledge base ConceptNet (Ghosal et al., 2020) is used as the reference. In the third module, we construct the multi-source semantic graph to model the various semantic relations residing in the three semantic sources, and adopt GCN to fulfil the sarcasm reasoning. In the last module, we generate the target sarcasm explanation with the BART (Lewis et al., 2020) decoder based on the three semantic sources. We conduct extensive experiments on a public released multimodal sarcasm explanation dataset, on which our method outperforms the best baseline by 28.90 and 22.47 in terms of BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004), respectively.

Our contributions can be concluded as follows.

- We propose a novel mulTi-source sEmantic grAph-based Multimodal sarcasm explanation scheme, where the fine-grained semantic information of the visual modality and the external knowledge concepts are jointly incorporated.

- As far as we know, we are the first to adopt the object-level metadata of the visual modality to promote the multimodal sarcasm explanation generation by the generative pre-trained language model.

- We propose a multi-source semantic graph, which is able to comprehensively capture the semantic relation among the input caption, input image, and external knowledge concepts. As a byproduct, we release our code and parameters[1] to facilitate this community.

2 Related Work

Our work is related to sarcasm detection and sarcasm-related generation.

2.1 Sarcasm Detection

Sarcasm detection aims to detect whether a post contains the sarcasm meaning. Early studies on sarcasm detection (Bouazizi and Ohtsuki, 2016; Felbo et al., 2017) mainly use hand-crafted features, such as punctuation marks, POS tags, emojis, and lexicons, to detect the sarcastic intention. Later, with the development of deep learning techniques, some researchers resorted to neural network architectures for sarcasm detection (Tay et al., 2018; Babanejad et al., 2020). Although these efforts have achieved promising progress, they focused on the text-based sarcasm detection, overlooking that the multimodal posts have been popping up all over the internet. Therefore, Schifanella et al. firstly proposed the multimodal sarcasm detection task and introduced a framework that fuses

[1] https://github.com/LiqiangJing/TEAM.

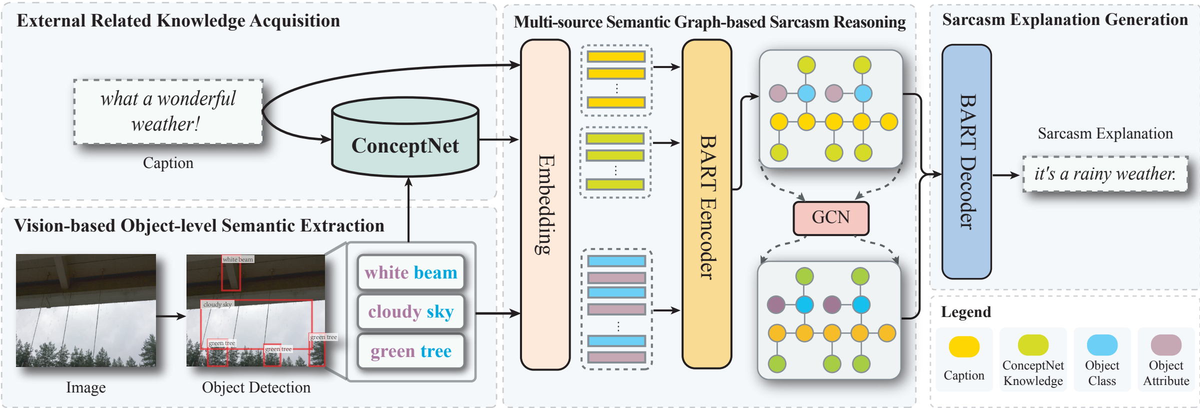

Figure 2: The architecture of the proposed TEAM, which consists of four key components: Vision-based Object-level Semantic Extraction, External Related Knowledge Acquisition, Multi-source Semantic Graph-based Sarcasm Reasoning, and Sarcasm Explanation Generation.

the textual and visual information with Convolutional Neural Networks (Ma et al., 2015) to detect the sarcasm intention. One limitation of this work is that it ignored the fine-grained ironic semantic relation in the multimodal input. Consequently, to boost the model performance, the following research efforts (Qiao et al., 2023; Kumar et al., 2022; Chakrabarty et al., 2020) resort to the Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) to mine inter-modal and intra-modal semantic association. Nevertheless, these efforts can only recognize whether a multimodal post contains the sarcastic meaning, but cannot explain why it is sarcastic, which is also important for various applications (Desai et al., 2022).

2.2 Sarcasm-related Generation

Apart from sarcasm detection, a few efforts attempted to conduct the sarcasm analysis by generating natural language. For example, some studies (Peled and Reichart, 2017; Dubey et al., 2019) resorted to machine translation models to generate non-sarcastic interpretation for the sarcastic text, which can help the smart customer service understand users' sarcastic comments and posts on various platforms. In addition, Mishra et al. employed unsupervised methods to transform a negative sentiment sentence to a sarcastic text in the context of dialog systems, which can make the agent's responses more natural and attractive to the user. Notably, these methods also only focus on text-based generation. Beyond them, recently, Desai et al. first proposed the multimodal sarcasm explanation task to support the sarcasm analysis and released a dataset, whose explanations are manually annotated. This method adopts the generative language model BART as the backbone, where the the global visual feature of the input image is incorporated with a cross-modal attention mechanism. Despite its remarkable performance, this method overlooks the gap between the visual feature space and the BART decoder semantic space, the object-level metadata of the image, and the potential external knowledge, which are the major concerns of our model.

3 Task Formulation

Suppose we have a training dataset $\mathcal{D}$ composed of N samples, *i.e.,* $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$. Each sample $d_i = \{T_i, V_i, Y_i\}$, where $T_i = \{t_1^i, t_2^i, \cdots t_{N_{t_i}}^i\}$ denotes the input caption which contains N_{t_i} tokens, V_i is the input image, and $Y_i = \{y_1^i, y_2^i, \cdots y_{N_{y_i}}^i\}$ denotes the target explanation text consisting of N_{y_i} tokens. Notably, N_{t_i} and N_{y_i} vary on different samples. Based on these training samples, our target is to learn a multimodal sarcasm explanation model $\mathcal{F}$ that is able to generate the sarcasm explanation based on the given multimodal input as follows,

$$\hat{Y}_i = \mathcal{F}(T_i, V_i|\Theta) \tag{1}$$

where Θ is a set of to-be-learned parameters of the model $\mathcal{F}$. $\hat{Y}_i$ is the generated explanation text by $\mathcal{F}$. For simplicity, we temporally omit the subscript i that indexes the training samples.

4 Method

In this section, we detail the four components of the proposed TEAM, as shown in Figure 2.

4.1 Vision-based Object-level Semantic Extraction

Considering that only the key visual information (*i.e.,* the objects in images) can demonstrate the sarcasm semantic, we propose to extract the object-level features of the image. Specifically, we feed the image into the Faster-RCNN (Anderson et al., 2018). Then for each region, it outputs not only the visual features (*e.g.,* content feature and positional feature) but also certain textual labels (*e.g.,* object class and object attribute). In our context, we only adopt the textual output, since we believe that textual labels contain rich semantics regarding the object, which should be beneficial towards the sarcasm reasoning, and fit better with the following encoding of the BART. Moreover, to ensure the quality of extracted object-level semantics, we only keep the top K regions with the highest confidence. Accordingly, for each image, we can obtain K objects, each of which is associated with a class name and an attribute value. Formally, we have,

$$\{(o_1, a_1), \cdots, (o_K, a_K)\} = \text{F-RCNN}(V) \quad (2)$$

where o_j and a_j are the extracted object class and attribute of the j-th object, respectively.

4.2 External Related Knowledge Acquisition

As aforementioned, the knowledge inferred by the input caption can support the sarcasm explanation generation since it may supply some concepts that appeared in the explanation or help the ironic semantic understanding with some sentiment knowledge. Specifically, we choose ConceptNet that describes general human knowledge in graph format[2] as the source of external knowledge, which involves 3.1 million concepts, and 38 million relations. Given our context of sarcasm explanation generation, we adopt the preprocessed ConceptNet (Li et al., 2022) that particularly covers the commonsense knowledge and emotional lexical knowledge, which plays an important role in the sarcasm reasoning.

To acquire the related external knowledge for the given multimodal input, *i.e.,* (T, V), we first identify all the concepts in ConceptNet that are mentioned in the input caption and the object meta-data (*i.e.,* object class and object attribute) derived by Faster-RCNN. Let $\{c_1, \cdots, c_{N_c}\}$ be the set of identified concepts, where N_c is the total number of identified concepts. We then use these identified concepts as the anchors to obtain the related concepts as the external knowledge for the multimodal input. Specifically, for each anchor concept e, we retrieve all its one-hop neighboring concepts from the knowledge graph ConceptNet and deem them as the external knowledge for c. Mathematically, let $\mathcal{N}(c)$ be the set of neighboring concepts of the concept c in ConceptNet. Then the related external knowledge for the multimodal input can be represented as $\{\mathcal{N}_{c_1}, \mathcal{N}_{c_2}, \cdots, \mathcal{N}_{c_{N_c}}\}$.

4.3 Multi-source Semantic Graph-based Sarcasm Reasoning

By now, we have three kinds of semantic sources: original input caption, object textual meta-data extracted from the input image, and external related textual concepts. To extract their features, we resort to the BART encoder, which has achieved compelling success on various natural language processing tasks, such as sentiment analysis (Mahdaouy et al., 2021) and multimodal summarization (Xing et al., 2021). Since the three semantic sources share the same token form, we first concatenate them into a sequence of tokens, denoted as X, and then feed X into the BART encoder $\mathcal{E}$ as follows,

$$\mathbf{H} = \mathcal{E}(X), \quad (3)$$

where $\mathbf{H} \in \mathbb{R}^{N \times D}$ is the encoded representation matrix, each column of which corresponds to a token, and N is the total number of tokens in X.

In fact, there are rich semantic relations resided in the three kinds of semantic sources that can be used for the sarcasm reasoning and the corresponding explanation generation. For example, the semantic correlation among tokens in the input caption can help the intra-modal inconsistency mining; the semantic correspondence between tokens in the input caption and that in the object meta-data can facilitate the cross-modal inconsistency uncovering. Moreover, linking the retrieved knowledge concepts to tokens in the input caption as well as those in the object meta-data promotes the semantic understanding of the multimodal input.

In light of this, for each sample d, we propose to construct a multi-source semantic graph $\mathcal{G}$ to comprehensively capture the above semantic relations. Let $\mathcal{H} = \{h_1, \cdots, h_N\}$ denote the set of nodes, which correspond to N tokens in X and can be divided into three categories: textual caption nodes, object nodes, and knowledge nodes. The

[2] https://conceptnet.io/.

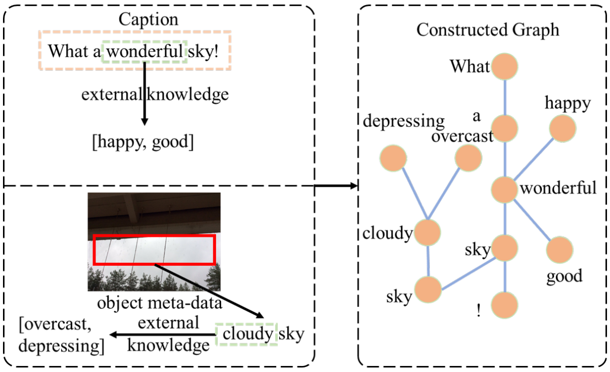

Figure 3: An example of the multi-source semantic graph construction process.

representations of these nodes are initialized by $\mathbf{H}$.

The edges of this graph are defined according to the semantic relations among these nodes as follows. 1) We first link the semantically correlated text nodes by adding an edge between each pair of adjacent tokens in the input caption. 2) We then introduce an edge between each object class and its corresponding object attribute, to link the object nodes that characterize the same object. 3) To capture the cross-modal semantic relation, we build an edge between each object class and its most similar token in the input caption, where the cosine similarity metric is used. And 4) for each retrieved knowledge concept, we link it with tokens in the input caption and object meta-data that act as the anchor concept in the aforementioned knowledge concept retrieval process. Formally, let $\mathbf{A} \in \mathbb{R}^{N \times N}$ denote the adjacency matrix of our constructed multi-source semantic graph. In order to facilitate understanding, we describe the construction process of the multi-source semantic graph in Figure 3.

Thereafter, we resort to the commonly used GCNs to conduct the sarcasm reasoning. Specifically, suppose we adopt L layers of GCN. Then all the node representations are iteratively updated as follows,

$$\mathbf{G}_l = ReLU(\tilde{\mathbf{A}}\mathbf{G}_{l-1}\mathbf{W}_l), l \in [1, L], \tag{4}$$

where $\tilde{\mathbf{A}} = (\mathbf{D})^{-\frac{1}{2}}\mathbf{A}(\mathbf{D})^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, and $\mathbf{D}$ is the degree matrix of $\mathbf{A}$. In addition, $\mathbf{W}_l \in \mathbb{R}^{D \times D}$ is a trainable parameter of the l-th GCN layer. $\mathbf{G}_l$ are the representations of nodes obtained in the l-th layer GCN, where $\mathbf{G}_0 = \mathbf{H}$ is the initial node representation.

Table 1: Statistics of the MORE dataset. Avg.length and $|\mathcal{V}|$ denote the average length of text and the vocabulary size, respectively.

Name	#Samples	Caption		Explanation	
		Avg.length	$\lvert\mathcal{V}\rvert$	Avg.length	$\lvert\mathcal{V}\rvert$
Train	2,983	19.75	9,677	15.47	5,972
Val	175	18.85	1,230	15.39	922
Test	352	19.43	2,172	15.08	1,527
Total	3,510	19.68	10,865	15.43	6,669

4.4 Sarcasm Explanation Generation

The final nodes representation $\mathbf{G}_L$ obtained by the L-layer GCN should absorb rich semantic information from their correlated nodes and can be used as the input for the following sarcasm explanation generation. Considering that the residual connection always performs well in the task of text generation (Vaswani et al., 2017), we also introduce a residual connection for generating the sarcasm explanation. Specifically, we first fuse the initial and final nodes representations as follows,

$$\mathbf{R} = \mathbf{H} + \mathbf{G}_L \tag{5}$$

where $\mathbf{R} \in \mathbb{R}^{N \times D}$ denotes the fused node representation. We then feed $\mathbf{R}$ to the decoder of the pre-trained BART. The decoder works in an auto-regressive manner, namely, producing the next word by considering all the previously decoded outputs as follows,

$$\hat{\mathbf{y}}_t = BART\_Decoder(\mathbf{R}, \hat{Y}_{<t}), \tag{6}$$

where $t \in [1, N_y]$ and $\hat{\mathbf{y}}_t \in \mathbb{R}^{|\mathcal{V}|}$ is the predicted t-th token's probability distribution of the target sarcasm explanation. $\hat{Y}_{<t}$ refers to the previously predicted t-1 tokens. Notably, in the training phase, to avoid the accumulated error, $\hat{Y}_{<t}$ will be replaced by $Y_{<t}$, *i.e.,* the previous $t-1$ tokens in the target sarcasm explanation.

For optimizing our TEAM, we adopt the standard cross-entropy loss function as follows,

$$\mathcal{L}_{Gen} = -1/N_y \sum_{i=1}^{N_y} \log(\hat{\mathbf{y}}_i[t]), \tag{7}$$

where $\hat{\mathbf{y}}_i[t]$ is the element of $\hat{\mathbf{y}}_i$ that corresponds to the i-th token of the target explanation, and N_y is the total number of tokens in the target sarcasm explanation Y.

5 Experiment

5.1 Dataset

We conducted experiments on the multimodal sarcasm explanation dataset **MORE** (Desai et al.,

Table 2: Comparing the generation performance of our model against state-of-the-art baselines on the MORE dataset. The best results are in boldface, while the second best are underlined.

Model	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
PGN	17.54	6.31	2.33	1.67	16.00	17.35	6.90	15.06	84.80	85.10	84.90	49.42
Transformer	11.44	4.79	1.68	0.73	15.90	17.78	5.83	9.74	83.40	84.90	84.10	52.55
MFFG-RNN	14.16	6.10	2.31	1.12	16.21	17.47	5.53	12.31	81.50	84.00	82.70	44.65
MFFG-Transf	13.55	4.95	2.00	0.76	15.14	16.84	4.30	10.97	81.10	83.80	82.40	41.58
M-Transf	14.37	6.48	2.94	1.57	18.77	20.99	6.98	12.84	86.30	86.20	86.20	53.85
ExMore	19.26	11.21	6.56	4.26	25.23	27.55	12.49	19.16	88.30	87.50	87.90	59.12
TEAM-w/o-Know	52.63	42.42	35.80	30.91	48.67	49.28	33.18	48.53	90.90	91.40	91.10	71.58
TEAM	**55.32**	**45.12**	**38.27**	**33.16**	**50.58**	**51.72**	**34.96**	**50.95**	**91.80**	**91.60**	**91.70**	**72.92**

(a) All samples

Model	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
PGN	17.87	6.37	1.92	1.26	16.43	17.80	6.92	15.62	84.70	85.20	84.90	48.77
Transformer	11.65	5.65	1.73	0.69	16.16	17.41	6.26	10.13	83.60	85.10	84.30	48.40
MFFG-RNN	15.43	6.82	2.46	1.33	17.40	18.61	5.71	12.98	81.60	84.30	82.90	42.72
MFFG-Transf	13.28	5.35	1.49	0.26	14.90	16.80	4.35	11.19	81.30	84.00	82.60	41.68
M-Transf	14.91	6.90	2.66	0.83	19.34	21.05	7.08	13.91	86.50	86.30	86.40	51.77
ExMore	19.47	11.69	6.82	4.27	24.92	27.12	12.12	19.20	88.30	87.60	88.00	56.95
TEAM-w/o-Know	53.43	43.41	36.77	31.78	49.72	51.12	34.78	49.24	91.50	91.90	91.80	71.62
TEAM	**56.45**	**46.34**	**39.58**	**34.34**	**52.79**	**53.81**	**36.78**	**51.62**	**92.40**	**92.90**	**92.30**	**73.35**

(b) Non-OCR samples

Model	**BLEU**				**Rouge**			**METEOR**	**BERT-Score**			**Sent-BERT**
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
PGN	17.19	6.08	2.49	1.79	15.55	16.92	6.76	14,64	84.90	84.90	84.90	49.53
Transformer	10.68	4.01	1.49	0.71	15.04	17.25	5.32	8.99	83.20	84.70	83.90	53.94
MFFG-RNN	12.18	4.92	1.73	0.88	14.01	15.18	4.56	10.64	81.20	83.70	82.40	45.91
MFFG-Transf	12.87	4.12	1.69	0.62	14.20	15.54	3.53	9.70	81.00	83.60	82.30	41.13
M-Transf	14.06	6.25	3.22	2.28	18.42	21.04	7.01	12.06	86.20	86.10	86.10	55.66
ExMore	19.40	11.31	6.83	4.76	25.66	28.02	12.10	19.15	88.20	87.50	87.90	60.82
TEAM-w/o-Know	51.91	41.51	34.85	29.85	47.53	49.00	32.77	47.94	90.50	**91.00**	90.70	71.43
TEAM	**52.88**	**43.08**	**36.81**	**32.34**	**48.46**	**49.68**	**33.83**	**49.25**	**90.90**	90.00	**90.80**	**71.93**

(c) OCR samples

2022). It is created by collecting sarcastic posts from various social media sites (Twitter[3], Instagram[4] and Tumblr[5]), where the sarcasm explanation for each post is manually annotated. Finally, this dataset contains 3,510 triplets in the form of <$image, caption, explanation$>, including 2,983 for training, 175 for validation, and 352 for testing. Statistics of this dataset are summarized in Table 1.

5.2 Experimental Setup

We adopted the bart-base-chinese model provided by huggingface[6] as the backbone of our model. In practice, the total number of tokens in each sample, *i.e.,* N, is unified to 256 by padding or truncation operations. The feature dimension D is set to 768, and the largest number of objects we allow to extract from an image, *i.e.,* K, is set to 36. We used AdamW (Loshchilov and Hutter, 2017) as the optimizer and set the learning rate of GCN layers to 1e-3 and that of the BART to 1e-4. The batch size is set to 16 and the maximum number of epochs for model training is set to 20. Following the previous work (Desai et al., 2022), we employed BLEU-1, BLEU-2, BLEU-3, BLEU-4 (Papineni et al., 2002), ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERT-Score (Zhang et al., 2020) and Sent-BERT (Reimers and Gurevych, 2019) to evaluate the performance of text generation models.

[3]`https://twitter.com/home`.
[4]`https://www.instagram.com/`.
[5]`https://www.tumblr.com/`.
[6]`https://huggingface.co/facebook/bart-base`.

5.3 On Model Comparison

To validate our TEAM, we compared it with the following existing methods.

- **PGN** (See et al., 2017). Pointer Generator Network is a text-based generation model, which generates the text with not only a conventional decoder but also a copy mechanism that copies words directly from input caption.
- **Transformer** (Vaswani et al., 2017). This is also a text-based generation baseline, which generates the text with the advanced transformer architecture.
- **MFFG-RNN** and **MFFG-Trans**. These are two variations of MFFG (Liu et al., 2020), a multimodal-based generation model for video summarization, where MFFG-RNN and

Table 3: Experiment results of ablation study. The best results are in boldface.

Model	BLEU				Rouge			METEOR	BERT-Score			Sent-BERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
w/o-Caption	22.85	11.83	7.30	4.64	21.27	18.19	6.26	16.54	86.40	86.10	86.20	53.82
w/-Visual	49.97	39.45	32.76	27.78	46.12	46.34	30.21	40.86	90.10	89.70	89.90	67.02
w/o-Obj	53.89	43.18	36.65	31.86	49.13	50.48	34.53	49.38	90.80	91.20	91.00	72.27
w/o-Graph	53.39	42.90	36.08	31.65	48.17	50.25	34.21	49.21	91.40	89.70	90.50	71.77
w/-FullGraph	32.84	18.74	12.29	8.44	29.21	29.20	11.69	22.31	87.10	87.30	87.40	62.21
TEAM	**55.32**	**45.12**	**38.27**	**33.16**	**50.58**	**51.72**	**34.96**	**50.95**	**91.80**	**91.60**	**91.70**	**72.92**

(a) All samples

Model	BLEU				Rouge			METEOR	BERT-Score			Sent-BERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
w/o-Caption	23.31	12.53	8.23	5.02	22.23	19.09	7.83	17.42	87.50	87.30	87.40	54.97
w/-Visual	50.29	40.31	33.82	28.41	47.24	47.38	31.37	41.75	90.50	90.10	90.30	67.81
w/o-Obj	55.32	44.87	37.82	33.96	50.58	52.45	36.12	51.06	91.60	91.80	91.90	72.98
w/o-Graph	54.65	43.82	37.29	32.27	50.42	51.18	35.26	49.25	91.80	90.20	91.30	72.31
w/-FullGraph	33.56	19.35	13.62	9.18	30.87	30.22	13.04	23.21	87.20	87.40	87.50	63.92
TEAM	**56.45**	**46.34**	**39.58**	**34.34**	**52.79**	**53.81**	**36.78**	**51.62**	**92.40**	**92.90**	**92.30**	**73.35**

(b) Non-OCR samples

Model	BLEU				Rouge			METEOR	BERT-Score			Sent-BERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
w/o-Caption	21.35	10.23	6.21	3.25	20.57	16.61	5.02	15.83	85.20	85.10	85.40	52.94
w/-Visual	48.37	38.25	31.26	26.28	44.60	45.12	29.02	39.97	89.90	89.50	89.70	66.54
w/o-Obj	52.19	42.86	35.24	31.02	46.88	49.60	33.17	48.46	90.20	90.60	90.70	71.64
w/o-Graph	51.32	41.91	34.25	31.23	46.57	49.26	33.97	49.18	90.70	89.40	89.60	70.31
w/-FullGraph	32.13	18.12	11.46	7.76	28.16	28.35	10.16	21.45	86.80	87.10	87.30	60.57
TEAM	**52.88**	**43.08**	**36.81**	**32.34**	**48.46**	**49.68**	**33.83**	**49.25**	**90.90**	90.00	**90.80**	**71.93**

(c) OCR samples

MFFG-Trans adopt the RNN and transformer architecture as the decoder, respectively.

- **M-Transf** (Yao and Wan, 2020). To use the visual modality to improve the quality of multimodal machine translation, this model equips Transformer with the multimodal self-attention mechanism to avoid encoding irrelevant information in images.
- **ExMore** (Desai et al., 2022). This is the most relevant baseline, which is designed for the task of multimodal sarcasm explanation. This method adopts BART as the model backbone and employs the cross-modal attention to inject the visual information into BART.
- **TEAM-w/o-Know**. Considering that all the baselines do not use the external knowledge, for fair comparison, we also introduced this variant of our model, where all the knowledge concepts are removed from our model.

Following the existing work (Desai et al., 2022), we conducted the performance comparison among different methods under three dataset configurations: a) on all samples, b) only on Non-OCR samples, and c) only on OCR samples. OCR samples denote the samples whose images contain embedded texts, while Non-OCR samples do not. We reported the experiment results in Table 2. From this table, we have several observations. (1) Both our complete model TEAM and its variant TEAM-w/o-Know consistently exceed all the state-of-the-art baselines in terms of all the metrics across different dataset configurations, which thoroughly demonstrates the superiority of our model. (2) The multimodal-based generation models (*e.g.,* MFFG-RCNN and MFFG-Transf) do not always perform better than the text-based models (*e.g.,* PGN). This implies that the performance of the model could be worse if the visual modality is not used properly. 3) The performance of our model on Non-OCR samples is higher than that on OCR samples across all metrics. The possible reason is that since our model only considers the object-level meta-data, the embedded text in the image could be ignored, leading to the information loss. In spite of this, our model still achieves a significant improvement over the best baseline on the Non-OCR samples.

5.4 On Ablation Study

We introduced the following variants of our model for the ablation study. 1) **w/o-Caption**. To evaluate the role of the caption in sarcasm explanation generation, we did not utilize the caption in this model. 2) **w/-Visual**. To show the superiority of using the object meta-data over the object visual feature, we adopted the object visual features extracted by Vit (Dosovitskiy et al., 2021), and concatenated them with the textual caption features to derive **H**, while the object meta-data is totally removed. 3)

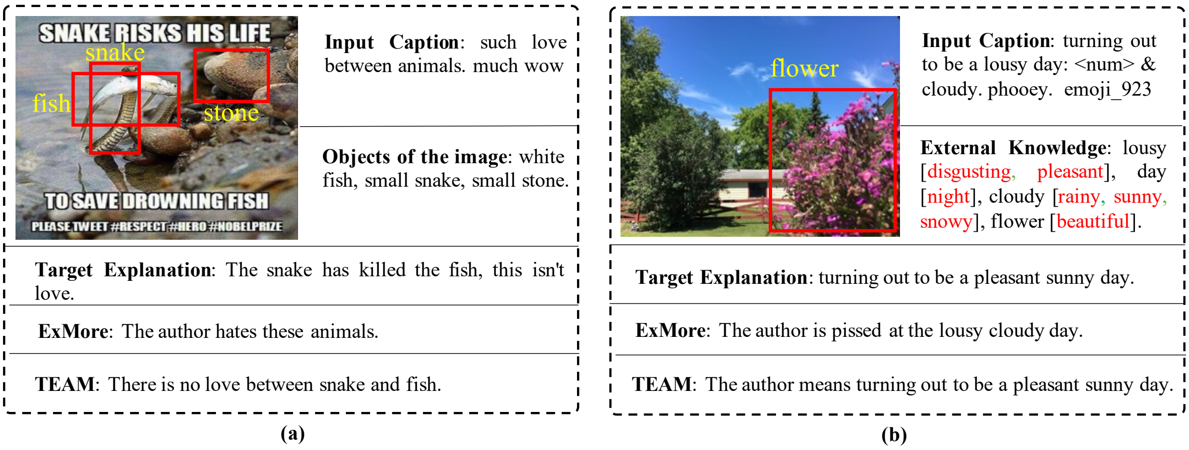

Figure 4: Comparison between the explanation generated by our model and the best baseline ExMore on two testing samples. The words in red are the related external knowledge concepts.

w/o-Obj. To show the benefit of extracting the key objects from the images, we omitted the object meta-data from the input. 4) **w/o-Graph**. To verify the necessity of building the multi-source semantic graph for sarcasm reasoning, we removed $\mathbf{G}_L$ and only fed $\mathbf{H}$ into the BART decoder. 5) **w/-FullGraph**. To further investigate the semantic relations of our multi-source semantic graph, we erased all the semantic relations and transformed the semantic graph to a full connection graph.

The ablation study results are shown in Table 3. From this table, we have the following observations. 1) w/o-Caption performs terribly compared with TEAM. This is reasonable since the caption is the main source for delivering the ironic intention. 2) TEAM exceeds w/-Visual. It demonstrates that the object-level metadata is better than the visual feature to stimulate the generation of sarcasm explanation with BART. 3) TEAM consistently outperforms w/o-Obj across different evaluation metrics. It confirms the necessity of using object-level feature for generating sarcasm explanation. 4) TEAM outperforms w/o-Graph, denoting that the graphs are essential to capture the ironic intention in the multimodal sarcastic posts. And 5) w/-FullGraph performs worse than TEAM, which verifies the utility of proposed semantic relations.

5.5 On Case Study

To get an intuitive understanding of how our model works on multi-modal sarcasm explanation, we showed two testing samples in Figure 4 due to the limited space. For comparison, we also displayed the explanation results of the best baseline ExMore. In case (a), as you can see, our model performs better than ExMore in terms of the quality of the generated sarcasm explanation. This may be attributed to the fact that our model considers the object-level metadata (*i.e.,* “fish” and “snake”) of the image, which benefits the sarcasm reasoning and explanation generation. In case (b), our model correctly explains the sarcasm, while ExMore failed. By analyzing the retrieved external knowledge concepts, we noticed that the concept “disgusting” benefits the semantic learning of the input caption, while concepts “sunny” and “beautiful” promotes the semantic interpretation of the input image. Moreover, the related concept “pleasant” of the word “lousy” contributes to the sarcasm explanation generation. Overall, these two cases intuitively show the benefits of incorporating both object-level meta-data and external knowledge concepts in the context of multimodal sarcasm explanation.

6 Conclusion and Future Work

In this work, we propose a novel multi-source semantic graph-based multimodal sarcasm explanation generation scheme. Experimental results on a public dataset demonstrate the superiority of our model over existing cutting-edge methods, and validate the advantage of utilizing the object-level meta-data over the global visual feature of the image as well as the benefit of incorporating the external knowledge in the context of multimodal sarcasm explanation. Particularly, we notice that our model performs worse on OCR samples than on Non-OCR samples. This is due to that our model currently ignores the text embedded in the image. In the future, we plan to incorporate the embedded text, which could indicate important clues for sarcasm explanation, to boost the model performance.

Acknowledgements

This work is supported by the Shandong Provincial Natural Science Foundation, No.:ZR2022YQ59; the National Natural Science Foundation of China, No.:62236003 and No.:62172261.

Limitations

Our work mainly suffers from two key limitations. 1) Ignore that the text embedded in the image could also reflect the sarcastic intention. As mentioned previously, we found that our model performs better on Non-OCR samples than the OCR samples. This may be due to the fact that our model ignores the text embedded in the image. Nevertheless, such embedded text could also indicate the ironic intention, (see Figure 3 (a)). We believe recognizing the text of the image can boost the performance of existing multimodal sarcasm explanation models. 2) Ignore that different knowledge concepts may contribute differently to the sarcasm reasoning. As shown in Figure 3 (b), the related concepts "disgusting" and "pleasant" should contribute more than the concept "night" in the sarcasm reasoning. Currently, our model equally treats all the knowledge concepts.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In *Proceedings of the Association for Computational Linguistics Student Research Workshop*, pages 107–113. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition*, pages 6077–6086. IEEE.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the International Conference on Computational Linguistics*, pages 225–243. International Committee on Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Mondher Bouazizi and Tomoaki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986. Association for Computational Linguistics.

Poorav Desai, Tanmoy Chakraborty, and Md. Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI Conference on Artificial Intelligence, Conference on Innovative Applications of Artificial Intelligence, The Symposium on Educational Advances in Artificial Intelligence*, pages 10563–10571. AAAI Press.

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*. OpenReview.net.

Abhijeet Dubey, Aditya Joshi, and Pushpak Bhattacharyya. 2019. Deep models for converting sarcastic utterances into their non sarcastic interpretation. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 289–292. ACM.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625. Association for Computational Linguistics.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210. Association for Computational Linguistics.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196. Association for Computational Linguistics.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on*

Natural Language Processing of the Asian Federation of Natural Language Processing, pages 757–762. Association for Computer Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. OpenReview.net.

Shivani Kumar, Atharva Kulkarni, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5956–5968. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI Conference on Artificial Intelligence, Conference on Innovative Applications of Artificial Intelligence, The Symposium on Educational Advances in Artificial Intelligence*, pages 10993–11001. AAAI Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1834–1845. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *International Conference on Computer Vision*, pages 2623–2631. IEEE.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in arabic language. In *Proceedings of the Arabic Natural Language Processing Workshop*, pages 334–339. Association for Computational Linguistics.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 6143–6153. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1690–1700. Association for Computational Linguistics.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 3980–3990. Association for Computational Linguistics.

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1010–1020. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: knowledge enhanced multimodal BART for visual commonsense generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 525–535. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. OpenReview.net.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Last section

☒ A2. Did you discuss any potential risks of your work?
This is no potential risk for our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract section

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Our model significantly surpasses other model.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.