Multi-modal Sarcasm Generation: Dataset and Solution

Wenye Zhao[1], Qingbao Huang[1,2,3*], Dongsheng Xu[1], Peizhi Zhao[1]
[1]School of Electrical Engineering, Guangxi University, Nanning, Guangxi, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
[3]Key Laboratory of Big Data and Intelligent Robot (SCUT), Ministry of Education
{2112391074, 2112391059, 2112391073}@st.gxu.edu.cn, qbhuang@gxu.edu.cn

Abstract

As an interesting and challenging task, sarcasm generation has attracted widespread attention. Although very recent studies have made promising progress, none of them considers generating a sarcastic description for a given image - as what people usually do on Twitter. In this paper, we present a Multi-modal Sarcasm Generation (MSG) task: Given an image with hashtags that provide the sarcastic target, MSG aims to generate sarcastic descriptions like humans. Compared with textual sarcasm generation, MSG is more challenging as it is difficult to accurately capture the key information from images, hashtags, and OCR tokens and exploit multi-modal incongruity to generate sarcastic descriptions. To support the research on MSG, we develop MuSG, a new dataset with 5000 images and related Twitter text. We also propose a multi-modal Transformer-based method as a solution to this MSG task. The input features are embedded in the common space and passed through the multi-modal Transformer layers to generate the sarcastic descriptions by the auto-regressive paradigm. Both automatic and manual evaluations demonstrate the superiority of our method. The dataset and code will be available at github.com/lukakupolida/MSG.

1 Introduction

Sarcasm is a type of emotional expression that indirectly expresses contempt, shows irritation, or demonstrates humor. As a typical task on sarcasm, Sarcasm Generation (SG) is proposed to generate a sarcastic message for a given literal input (Joshi et al., 2015), which can express a variety of communicative intent such as evoking humor and diminishing or enhancing critique (Burgers et al., 2012). It can impact many downstream applications such as personalized dialog systems (Cho et al., 2022)

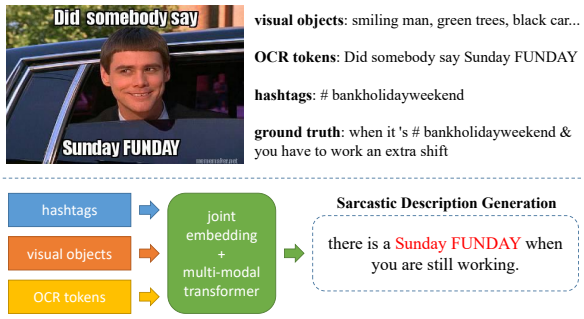


Figure 1: An example of Multi-modal Sarcasm Generation and the illustration of our proposed multi-modal Transformer-based architecture (MTMSG). We feed the features from hashtags, visual objects, and OCR tokens modalities into the multi-modal Transformer. Further, the sarcastic description is generated through iterative decoding with a pointer network and linear layers.

and news comment generation (Yang et al., 2019). Since SG is proposed, a surge of follow-up studies have been conducted (Peled and Reichart, 2017; Mishra et al., 2019; Chakrabarty et al., 2020; Oprea et al., 2021, 2022). Notably, the aforementioned SG studies have only been investigated in the textual field so far.

However, nowadays social platforms usually leverage multi-modal data where visual information is integrated with the text, making the analysis of uni-modal data in isolation and limitation. Therefore, research on multi-modal sarcasm is crucial and imperative. Studies on multi-modal sarcasm can be categorized into threefold: Multi-modal Sarcasm Detection (Cai et al., 2019; Pan et al., 2020; Xu et al., 2020; Liang et al., 2021, 2022), Multi-modal Target Identification (Wang et al., 2022), and Multi-modal Sarcasm Explanation (Kumar et al., 2022; Desai et al., 2022). Unfortunately, there has been no research touching Sarcasm Generation facing multi-modal information until now. We hope general artificial intelligence to learn creativity and associative skills, so learning to generate sarcastic descriptions towards multi-modal inputs like

*: Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2023, pages 5601–5613
July 9-14, 2023

humans deserves to study deeply.

Therefore, we propose a Multi-modal Sarcasm Generation task (MSG), which aims to generate sarcastic descriptions on social platforms like humans for a given image with the help of hashtags (cf. Figure 1). Compared with textual sarcasm generation, MSG is more challenging. Firstly, human expressions on social platforms are usually stylized with too many verbalized expressions like abbreviations and interjections. Secondly, accurately capturing information from the key visual regions which may contribute to the sarcasm remains a question. Finally, the incongruity between images and generated Sarcastic descriptions reflects human creativity, imagination, and associative skills, which are hard for machines to learn and construct.

To support the studies on MSG, we develop MuSG, a new dataset consisting of 5000 images and related sarcastic descriptions. We manually collect samples with clear sarcastic target from Twitter API and two existing multi-modal sarcasm detection datasets (Schifanella et al., 2016; Cai et al., 2019). The descriptions come with hashtags (the tokens with a '#' to indicate the topic of Twitter) that point the way to sarcasm generation (cf. Figure 1 *#bankholidayweekend*). The images contain OCR tokens information that can provide an associative context for sarcasm generation (cf. Figure 1 *Did somebody say Sunday FUNDAY*). With the well-formed dataset MuSG, researchers can easily conduct studies on MSG.

Consequently, as shown in Figure 1, we design a Multi-modal Transformer-based model (MTMSG) as a strong baseline for the proposed MSG task. Concretely, we first model spatial, semantic, and visual reasoning relations between multiple OCR tokens, hashtags, and visual features. Further, we map all the modality-specific features to the same reference and utilize the self-attention mechanism (Parikh et al., 2016) to capture the relationships between them. Finally, we combine the vocabulary with OCR tokens which usually contain sarcastic intent to capture the incongruity for sarcasm generations. With the ability to capture the intra- and inter-modality incongruity, our model is thus capable of effectively generating sarcastic descriptions.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to investigate the Multi-modal Sarcasm Generation task, which aims to generate sarcastic descriptions like humans for a given image with the help of hashtags.
- We develop MuSG, a new dataset consisting of 5000 image-text pairs for Multi-modal Sarcasm Generation. To our knowledge, it is the only dataset that can be applied to this task and evaluated automatically.
- We benchmark MuSG with a multi-modal Transformer-based model which can be served as a strong baseline.
- Empirical results show that our MTMSG outperforms all comparison models on all automatic evaluation metrics. We also perform extensive human evaluations to measure the Creativity, Sarcasticness, Coherence, and Image-Text Relation of generated descriptions.

2 Related Work

2.1 Textual Sarcasm Generation

Recently, sarcasm generation has attracted tremendous attention in the field of natural language processing. The studies can be roughly categorized into twofold: Joshi et al. (2015) and Oprea et al. (2021, 2022) generate sarcasm with a response generator, while Peled and Reichart (2017), Mishra et al. (2019), and Chakrabarty et al. (2020) generate sarcasm with a paraphrase generator. However, these studies concentrate only on the generation of sarcasm in the textual domain, till now there has been no relevant effort on Multi-modal Sarcasm Generation. Applying machines to think and further imagine like humans is a creative and challenging task, fulfilling our imagination for the future of general artificial intelligence. Accordingly, we strongly believe that our proposed MSG will lead to a deeper understanding and expression of individuals' intent on social media.

2.2 Research on Multi-modal Sarcasm

With the rapid development of mobile Internet, research on multi-modal sarcasm has come into focus. Schifanella et al. (2016) pioneer the multi-modal sarcasm and propose a dataset to collect 10000 sarcastic posts from Twitter, Instagram, and Tumblr. Cai et al. (2019) and Castro et al. (2019) extend the research and develop richer and better-formed datasets based on Twitter and conversational audio-visual utterances, respectively. Since then, a surge of studies have been conducted on multi-modal sarcasm, which can be roughly divided into three

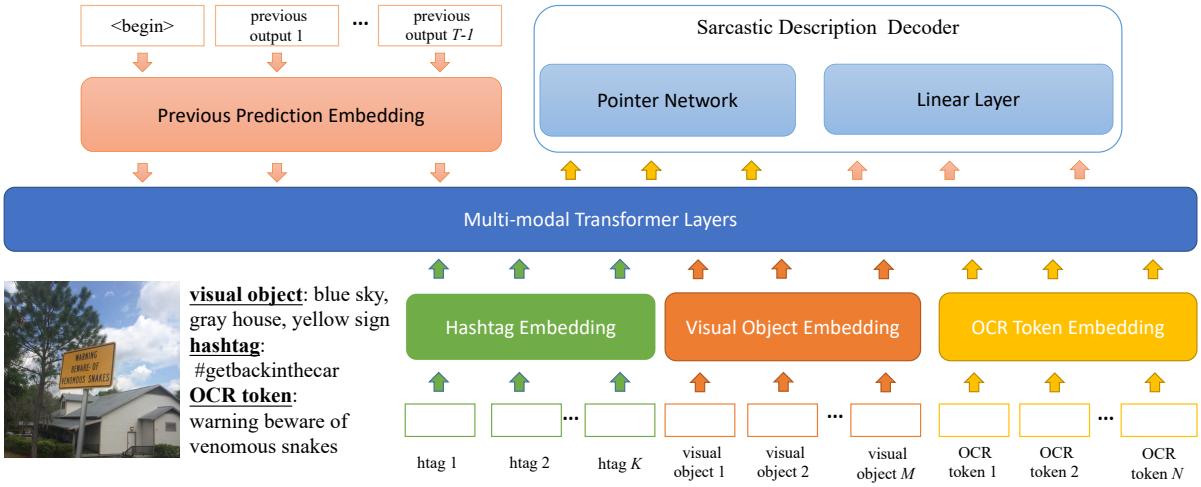


Figure 2: The architecture of our proposed MTMSG. The input features from hashtags, visual objects, and OCR tokens modalities are embedded in the common 768-dimensional reference space and passed through the multi-modal Transformer layers to generate the sarcasm sentence through the auto-regressive paradigm.

categories: Multi-modal Sarcasm Detection (Pan et al., 2020; Xu et al., 2020; Liang et al., 2021, 2022), which aims to detect whether the input sample is sarcastic; Multi-modal Sarcasm Target Identification (Wang et al., 2022), which aims to extract sarcasm targets from both texts and images; Multi-modal Sarcasm Explanation (Kumar et al., 2022; Desai et al., 2022), which aims to generate a natural language sentence to explain the intended irony in the sarcastic posts. However, there is no research considering Multi-modal Sarcasm Generation. Leveraging the multi-modal information to create sarcastic descriptions will increase the variety of responses for intelligent conversational agents and further serve downstream applications such as personalized dialog systems (Cho et al., 2022) and news comment generation (Yang et al., 2019). Therefore, it is crucial to study the Multi-modal Sarcasm Generation task.

3 Dataset and Metrics

In this section, we describe how the new dataset is constructed and how performance is evaluated.

3.1 Dataset

Since Twitter text contains varieties of sarcastic descriptions and intent detection in Twitter is a problem worth investigating, we focus on the Twitter-based dataset. We retrieve posts by querying hashtags to collect potential sarcastic samples. To create a well-formed and high-quality dataset **MuSG** for the MSG task, we collect publicly available Twitter posts using Twitter API and two existing multi-modal sarcasm detection datasets (Schifanella et al., 2016; Cai et al., 2019) to obtain 5000 samples that have clear sarcasm targets. For text data, we remove external links and mentions (@email); We remove strange and meaningless symbols such as the token *emoji-x* and other special tokens which are hard to understand (♠); We also remove text with more than 40 words, because if the text is too long, it is much more difficult to generate by machine with the same inputs. For image data, we remove text-based images (images consist of text only), images with number-based OCR tokens (OCR tokens in the images consist of numbers only), images with too many visual objects, and images with low resolution. The MuSG dataset is randomly split into 3536/723/741 (using 5:1:1 split) as Train/Valid/Test in the experiments.

Further, we conduct a comprehensive statistical analysis of our collected MuSG dataset as follows: First, the content categories of Twitter contain six major categories: politics, sports, games, dining, life, and others, with politics and life accounting for the largest share, reaching half of the total (cf. Table 1). Second, we count the subjects of sarcastic sources, of which 27.3% originate from both images and hashtags, 32.4% from hashtags only, and 40.3% from images only, so we can conclude that most of the sarcasm can be generated with the help of images and hashtags (cf. Table 2). Third, we count the size of the sarcasm targets, and we find that 25.8% of the sarcasm targets are small targets, making the MSG task more challenging (cf. Table 3). Finally, we also count the sentence style of the

ground truth, in which 56.3% are declarative sentences, 40.9% are imperative sentences, and only 2.8% are interrogative sentences or other sentences, which represents that sarcastic descriptions express a definite emotion in most cases (cf. Table 4).

Politics	Sports	Games	Dining	Life	Others
26.4%	8.7%	5.4%	13.4%	28.4%	17.7%

Table 1: The content categories of MuSG dataset.

Both in images and htags	Only in images	Only in htags
27.3%	40.3%	32.4%

Table 2: Statistics of the subject of the MuSG dataset.

Small	Medium	Large
1290 (25.8%)	1865 (37.3%)	1845 (36.9%)

Table 3: The size of the subject of the MuSG dataset.

Declarative	Imperative	Interrogative (Others)
56.3%	40.9%	2.8%

Table 4: The sentence types of the MuSG dataset.

3.2 Evaluation Metrics

We evaluate the generated descriptions both quantitatively (with standard automatic evaluation metrics) and qualitatively (with human evaluation metrics). For automatic evaluation metrics, we apply the Microsoft coco caption evaluation, which includes **BLEU** (**B1**, **B2**, **B3**, and **B4**) (Papineni et al., 2002), **METEOR** (Denkowski and Lavie, 2014), **ROUGE** (**R1**, **R2**, and **R_L**) (Lin, 2004), **CIDEr-D** (Vedantam et al., 2015), and **SPICE** (Anderson et al., 2016). For human evaluation metrics, we propose a set of 4 criteria to evaluate the generated descriptions: **1)Creativity** ("How creative are the generated descriptions?"), to judge if the generated descriptions are novel and attractive; **2)Sarcasticness** ("How sarcastic are the generated descriptions?"), to judge the degree of sarcasm (including irony and humor); **3)Coherence** ("How coherent are the generated descriptions?"), to judge if the generated descriptions are fluent and further easy to understand; **4)Image-text Relation** ("How relevant are the images and the generated descriptions?"), to judge if the generated descriptions are highly correlated with the given images.

4 Methodology

In this section, we describe our proposed **MTMSG**, a Multi-modal Transformer-based model for MSG. The input to this task is an image and the hashtags of the corresponding Twitter text, while the output is the generated descriptions that need to compare with the original Twitter text (Ground Truth). Yet, demonstrated by Pan et al. (2020), the information of OCR text usually provides the context of sarcasm, which may contribute to sarcasm generation. Therefore, we leverage the information from visual objects, hashtags, and OCR tokens for MSG.

The architecture of the proposed MTMSG is illustrated in Figure 2. Specifically, we first embed the three modalities in the same reference, and then feed them into a multi-modal Transformer to achieve intra- and inter-modality interactions. Finally, our models learn to generate sarcastic descriptions through iterative decoding with the help of a dynamic pointer network. In the decoding process, we leverage the previous output to predict the next generated word in an auto-regressive manner.

4.1 Uni-modal Feature Embedding

4.1.1 Hasgtag Embedding

Given a hashtag as a sequence of K words, we utilize FastText (Bojanowski et al., 2017) as the feature extractor to get the 300-dimensional vector x_k^{ft} $(k = 1, \cdots, K)$, which is a word embedding with sub-word information. Finally, we project the vector to a 768-dimensional semantic space to make sure the features from different modalities are embedded in the same reference. The final hashtag embedding x_k^{htag} is obtained by:

$$x_k^{htag} = LN(W_1 x_k^{ft}), \quad (1)$$

where W_1 is the learnable parameter and LN denotes layer normalization.

4.1.2 Visual Object Embedding

Given an image, we apply pretrained Faster R-CNN (Ren et al., 2015) as the detector to obtain the appearance feature x_m^{fr} of m-th visual object. Further, to leverage the spatial information of each object, we investigate a 4-dimensional location feature by $x_m^b = [x_{min}/W, y_{min}/H, x_{max}/W, y_{max}/H]$. Then we can obtain a list of 768-dimensional vectors x_m^{obj} as follows:

$$x_m^{obj} = LN(W_2 x_m^{fr}) + LN(W_3 x_m^b), \quad (2)$$

where W_2 and W_3 are learnable parameters, and LN denotes layer normalization.

4.1.3 OCR Token Embedding

For OCR token embedding, following the M4C-Captioner (Sidorov et al., 2020), to get a rich representation of OCR tokens, we leverage FastText (Bojanowski et al., 2017), Faster R-CNN (Ren et al., 2015), PHOC (Almazán et al., 2014) as the feature extractors to get sub-word feature x^{ft}, appearance feature x^{fr}, and character-level feature x^{p}, respectively. Given a set of N OCR tokens, the location feature of the n-th token is represented as $x_n^b = [x_{min}/W, y_{min}/H, x_{max}/W, y_{max}/H]$. Then the final OCR token embedding x_n^{ocr} is projected to a 768-dimensional vector as:

$$x_n^{ocr} = LN(W_4 x_n^{ft} + W_5 x_n^{fr} + W_6 x_n^{p}) + LN(W_7 x_n^{b}), \quad (3)$$

where W_4, W_5, W_6, and W_7 are learnable parameters, and LN denotes layer normalization.

4.2 Multi-modal Transformer

After extracting uni-modal feature embedding from three modalities, we concatenate the features and feed them into the multi-modal Transformer. In the multi-modal Transformer, the features fully interact to exploit intra- and inter-modality incongruity. Besides, the previous step of decoding output x_{t-1}^{dec} is also embedded and fed into the Transformers. Finally, the multi-modal Transformers obtain feature vectors as output: $[z^{htag}, z^{obj}, z^{ocr}, z_{t-1}^{dec}] = MMT([x^{htag}, x^{obj}, x^{ocr}, x_{t-1}^{dec}])$, where MMT denotes multi-modal Transformer.

4.3 Sentence Decoder

The sentence decoder takes the feature embedding output of multi-modal Transformers as input, predicts the score for each word, and selects the predicted word of each time step. We generate sarcastic descriptions through the auto-regressive paradigm. Remarkably, the OCR tokens detected in images usually contain intent information for capturing the multi-modal incongruity, while they are usually not involved in the common word vocabulary, so it is inappropriate to make predictions based on the fixed vocabulary. Therefore, we adopt different classifiers for common vocabulary and OCR tokens as follows:

$$y^t = argmax(f(z_{t-1}^{dec}), f_{DPN}(z_{t-1}^{dec}, z^{ocr})), \quad (4)$$

where f indicates the linear classifier for common vocabulary, and f_{DPN} indicates the dynamic pointer network(Vinyals et al., 2015). The captioning loss is computed by:

$$\mathcal{L} = -log\sum_{i=0}^{T}(P(y_t|(y_{0:T-1}, z^{htag}, z^{obj}, z^{ocr}))), \quad (5)$$

where $y_{0:T-1}$ denotes the generated sequence.

5 Experiments

5.1 Comparison Models

Due to the multi-modal nature of the input corpus, we compare our proposed MTMSG model with three categories of strong models adapted for the MSG task as follows:

1)Image-modality models: These models only leverage the visual features to generate sarcastic description, including **ViT** (Dosovitskiy et al., 2020), a powerful visual Transformer (with BART as the decoder); and **BLIP** (Li et al., 2022), a pre-trained image captioning model finetuned on this task.

2)Text-modality models: These models only leverage the hashtags and OCR tokens to generate sarcastic descriptions, including **Transformer** (Vaswani et al., 2017) and **Chandler** (Oprea et al., 2021), a very recent effort that generates sarcastic response to a given textual utterance.

3)Multi-modality models: These models utilize the information from images, hashtags, and OCR tokens for MSG, including **MFFG** (Liu et al., 2020), a multi-stage fusion mechanism with a forget fusion gate (both RNN and Transformer variants of MFFG); and **MMT** (Tang et al., 2022), a multi-modal Transformer for multi-modal learning.

5.2 Experiment Settings

For visual objects, we extract 100 object appearance features with the dimension 2048. Besides, we apply Google OCR API to detect sufficient OCR tokens with bounding boxes. The number of OCR tokens for each image is limited to 30 at most. We set the layer number of the multi-modal Transformer to 4 and the number of self-attention heads to 12. Following BERT-base (Devlin et al., 2019), we adopt default settings for other parameters. For MSG, we tokenize the text on whitespace and filter the special symbols that the model cannot recognize. The fixed common vocabulary has 11554 words. Furthermore, we train the model for 160 epochs on a single 3090Ti GPU, and the batch size is set to 64. We adopt the Adam optimizer, the initial learning rate is 1e-4 and declined to 0.1 times every 50 epochs. We monitor the CIDEr-D

Modality	Method	BLEU				Rouge			METEOR	CIDEr-D
		B1	B2	B3	B4	R1	R2	R_L		
image	ViT (ICLR 2020)	10.57	3.90	1.38	0.60	16.24	4.31	14.03	8.86	18.5
	BLIP (ICML 2022)	12.07	5.01	1.69	0.84	15.17	4.55	14.00	10.63	22.8
text	Transformer (NeurIps 2017)	11.33	4.68	1.57	0.62	17.67	5.72	15.79	9.63	24.8
	Chandler (EMNLP 2021)	17.19	7.82	4.62	2.91	18.72	7.01	16.41	10.30	28.8
image+text	MFFG-RNN (EMNLP 2020)	14.73	6.49	2.58	1.33	15.81	5.92	14.88	11.23	28.4
	MFFG-Transf (EMNLP 2020)	15.32	6.71	2.35	1.22	16.61	5.71	15.40	10.98	27.6
	MMT (IJCAI 2022)	18.62	7.64	4.04	2.64	22.02	8.00	19.77	12.95	35.2
	MTMSG(ours)	**21.37***	**13.32***	**8.67***	**6.37***	**26.44***	**11.38***	**24.12***	**16.05***	**48.6***

Table 5: Main experimental results regarding uni-modal and multi-modal scenarios. The best scores of each group are in bold. Results with * denote the significance tests of our MTMSG over the baseline models at p-value<0.05.

metric to choose the best model and evaluate it on the test set. Finally, we average the experimental results of our MTMSG over ten runs to ensure the statistical stability of the experimental results.

6 Experimental Results and Analysis

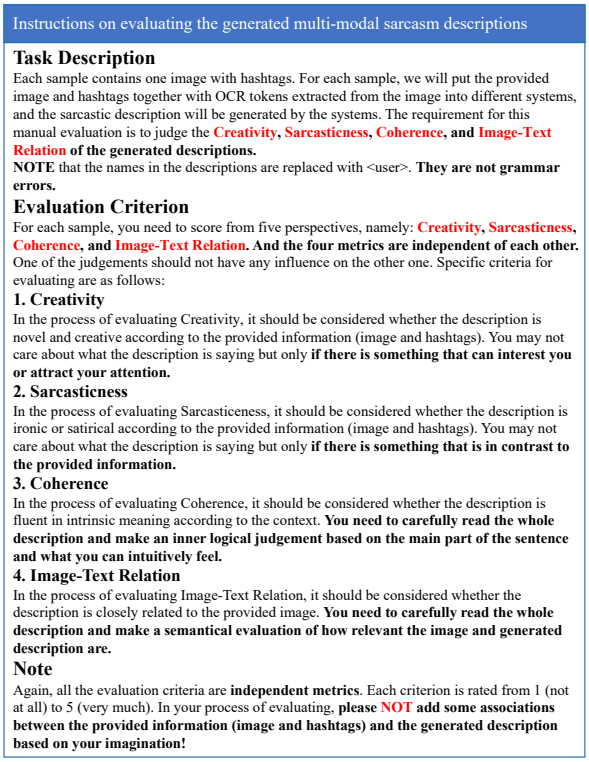
6.1 Main Results

Table 5 presents the comparative generation performances on the dataset. The experimental results illustrate that our model achieves the best performance across all of the competing strong baselines adopted for this task. Specifically, our model obtains BLEU scores of 21.37 (+2.75), 13.32 (+5.68), 8.67 (+4.63), and 6.37 (+3.73) on B1, B2, B3, and B4, respectively. Similarly, we exceed 4.42, 3.38, and 4.35 Rouge points at 26.44, 11.38, and 24.12 on R1, R2, and R_L, respectively. Our model also achieves improved performance on METEOR 16.05 (+3.1), and CIDEr-D 48.6 (+13.4).

Moreover, we can draw the following conclusions: 1) Notably, our proposed MTMSG outperforms existing strong baselines on all of the evaluation metrics (Significance Test, all p-value<0.05), which demonstrates the effectiveness of our proposed multi-modal Transformer-based model for MSG. 2) Models based on text modality perform better than the models based on image modality, which indicates more sarcastic information lies in the hashtags and OCR tokens. 3) Multi-modal models achieve much better performance than the uni-modal baselines overall, which implies that leveraging the information from images, hashtags, and OCR tokens is efficacious for MSG.

6.2 Human Evaluation

We also perform the human evaluation for assessing the quality of the MSG. We randomly select 200 samples from the test set. Given the provided information (image, hashtags, and OCR) together with the description generated by ours and the other 4 strong baselines (BLIP, Chandler, MFFG, MMT). Each criterion is rated from 1 (*not at all*) to 5 (*very*). We employ 5 evaluators to independently score the generated sarcastic descriptions from the five methods and the ground truth. Figure 3 shows the instructions released to the evaluators.


Instructions on evaluating the generated multi-modal sarcasm descriptions

Task Description

Each sample contains one image with hashtags. For each sample, we will put the provided image and hashtags together with OCR tokens extracted from the image into different systems, and the sarcastic description will be generated by the systems. The requirement for this manual evaluation is to judge the **Creativity**, **Sarcasticness**, **Coherence**, **and Image-Text Relation** **of the generated descriptions.**

NOTE that the names in the descriptions are replaced with <user>. **They are not grammar errors.**

Evaluation Criterion

For each sample, you need to score from five perspectives, namely: **Creativity**, **Sarcasticness**, **Coherence**, **and Image-Text Relation**. **And the four metrics are independent of each other.** One of the judgements should not have any influence on the other one. Specific criteria for evaluating are as follows:

1. Creativity

In the process of evaluating Creativity, it should be considered whether the description is novel and creative according to the provided information (image and hashtags). You may not care about what the description is saying but only **if there is something that can interest you or attract your attention.**

2. Sarcasticness

In the process of evaluating Sarcasticeness, it should be considered whether the description is ironic or satirical according to the provided information (image and hashtags). You may not care about what the description is saying but only **if there is something that is in contrast to the provided information.**

3. Coherence

In the process of evaluating Coherence, it should be considered whether the description is fluent in intrinsic meaning according to the context. **You need to carefully read the whole description and make an inner logical judgement based on the main part of the sentence and what you can intuitively feel.**

4. Image-Text Relation

In the process of evaluating Image-Text Relation, it should be considered whether the description is closely related to the provided image. **You need to carefully read the whole description and make a semantical evaluation of how relevant the image and generated description are.**

Note

Again, all the evaluation criteria are **independent metrics**. Each criterion is rated from 1 (not at all) to 5 (very much). In your process of evaluating, **please NOT add some associations between the provided information (image and hashtags) and the generated description based on your imagination!**

Figure 3: Instructions for human evaluation.

The comparative results are shown in Table 6. To measure the inner-annotator agreement, we calculated Fleiss' kappa and all the results show fair agreement ($0.2 \leq \kappa \leq 0.4$). From the diagram, we can obtain the following conclusions: 1) BLIP is superior in the research of image caption, so this method works well on the metrics of Coherence

MODEL	Cre.	Sac.	Coh.	I-T Rel.
GroundTruth	3.71	3.85	4.12	4.06
BLIP	1.12	1.26	**4.32**	**4.48**
Chandler	2.11*	2.04*	3.55	2.76
MFFG-Transf	1.88	1.94	3.21	2.67
MMT	1.93	1.87	3.37	2.91
MTMSG	**2.57**	**2.78**	3.62*	3.02*

Table 6: Average scores for generated sarcasm from a set of 4 criteria as human evaluation. Cre., Sar., Coh., and I-T Rel. denote the human evaluation metrics Creativity, Sarcasticness, Coherence, and Image-Text Relation, respectively. The scale ranges from 1 (*not at all*) to 5 (*very*). The bolded data is the best result of the specific metric, yet the data with * denotes the 2nd ranked.

and Image-Text Relation. Notably, it works even better than the ground truth, which further demonstrates that the Twitter text posted by humans is creative and full of imagination. 2) For Coherence and Sarcasticness, our proposed MTMSG performs better than any other baselines, which demonstrates that our model meets the basic target of the MSG task. 3) Moreover, we can see that all the baselines adopted for this task show poor performance on the metrics of Creativity and Image-Text Relation, which illustrates that more studies on improving the quality of the generated descriptions are needed. 4) Overall, the results strengthen that our model is superior to all the other baselines. However, a significant gap in performance still remains between humans and machines, which demonstrates that our proposed MSG task is challenging and worth further in-depth research.

6.3 Ablation Study

MODEL	B4	R_L	METEOR	CIDEr-D
M4TSD	**6.37**	**24.12**	**16.05**	**48.6**
w/o visual	6.02	23.31	15.57	45.2
w/o OCR	5.24	22.08	14.51	42.3
w/o htag	4.95	22.14	14.22	40.9
w/o htag, visual	4.32	21.33	13.23	34.1
w/o OCR, visual	4.13	20.95	12.91	32.7
w/o htag, OCR	1.73	15.41	11.88	24.2
w/o iterative	3.87	18.19	12.15	33.6
w/o pointer	2.51	17.54	11.88	28.1

Table 7: Experimental results of the ablation study. w/o denotes without. htag, OCR, visual, pointer, and iterative denote hashtags, OCR tokens, visual objects, pointer network, and iterative decoding, respectively.

We conduct an ablation study on the effectiveness of the three input modalities. Table 7 gives the experimental results of the ablation study. From the perspective of input corpus, we can draw the following conclusions: 1) Note that the removal of hashtags (w/o htag) significantly degrades the performance, which verifies the significance of guiding the direction of sarcasm generation as hashtags indicate the topic of Twitter; 2) Since the context of sarcastic information resides in the OCR tokens, it is hard to understand the sarcastic intent without the OCR tokens. As a result, removing the OCR tokens (w/o OCR) also leads to considerable performance decline; 3) Besides, from the results of w/o visual, we can conclude that the visual features are also beneficial to MSG; 4) Moreover, from experimental results which only utilize one modality at a time to observe how much modality-specific information contributes to the generation, we can conclude that textual information plays a major role in sarcasm generation. More remarkably, concentrating on our models, we can get the conclusions as follows: 1) Dynamic pointer network can predict the copying score between the decoding output and each OCR token. Since the OCR tokens and the common word vocabulary are usually complementary, the removal of the pointer network (w/o pointer) leads to apparent performance degradation; 2) Finally, without leveraging the iterative decoding method (w/o iterative), we only decode for one step. The sharp performance degradation indicates the iterative decoding strategy can improve the quality of generated sarcastic descriptions.

6.4 Case Study

We present some examples to analyze the performance of our model. Specifically, in Figure 4 (a), we can find that without hashtags and OCR tokens, our model can only describe a man in a black car. The hashtag and OCR token provide the timing factor *Sunday* and *FUNDAY* may guide the intention of sarcasm, which demonstrates the information from hashtags and OCR tokens is necessary for sarcasm generation. Yet MMT simply concatenates the visual information with the timing factor *Sunday*, and the generated description seems not to be sarcastic. Even sometimes the hashtag is not so useful, like *RTX Austin* in Figure 4 (b), it just tells the place of the scene. We can also understand the main intent of the sarcasm from the OCR tokens (*behind schedule*). Associated with the image, we can understand that the flight is behind schedule, which can help generate a sarcastic description.

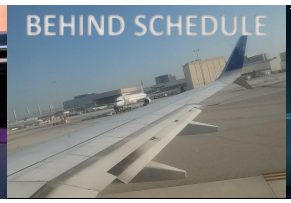
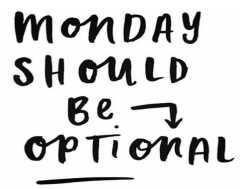
	(a)	(b)	(c)	(d)
Image				
hashtag	#bankholidayweekend	#rtxaustin	#speed	#mondaymorning
OCR token	Did somebody say Sunday FUNDAY?	BEHIND SCHEDULE	speed test; ping 255ms download 0.30mps;upload 0.28mps	Monday should be optional.
ground truth	when it's #bankholidayweekend & you have to work an extra shift.	flight delays are the best . # rtxaustin	here is my blazing fast unlimited <user> #speed . such great service for the price .	cntrl / alt / delete monday # mondaymorning # comedy # tired # coffeeaddict
MMT	a man in a black car on Sunday.	nice to see plane at the airport.	what a colorful speed table!	a paper with words concerning Monday.
our MTMSG	there is a Sunday FUNDAY when you are still working.	nice to see the flight was behind schedule.	here is so much for < user > speeds we must be amazing !	Optional Monday is the best thing about my life # comedy # nochill

Figure 4: Example of successfully generated cases. With the help of images, hashtags, and OCR tokens, our model is capable of generating sarcastic descriptions like humans to some degree.

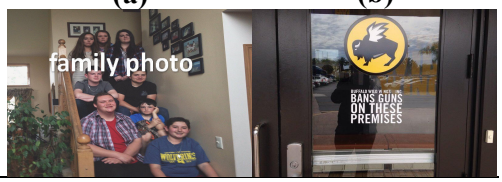
	(a)	(b)
Image		
hashtag	#family	#lunch
OCR token	family photo	bans guns on these premises
ground truth	can 't have #family without so many happy people	i felt so safe eating lunch today ! i 'm sure this sign kept all the bad guys out #lunch !
our model	can ' t have the more people !	i feel so excited about its finest lunch.

Figure 5: Example of falsely generated cases. In these cases, our provided hashtags and OCR tokens are not necessary for Multi-modal Sarcasm Generation.

While MMT still generates a literal description of the image (*plane at the airport*). Similarly, in Figure 4 (c), from the hashtag, we know that the sarcastic target is about speed, and from the OCR token, we can get that the intent is to express that speed is slow; in Figure 4 (d), from the hashtag, we know that the sarcastic target is about Monday morning, and from the OCR token, we can get that the intent is to express the hope of optional Monday. If we understand the intent, we can easily get better sarcastic descriptions instead of literal descriptions like MMT. All the examples demonstrate that with the help of proposed images, hashtags, and OCR tokens, our model is capable of generating sarcastic descriptions like humans to some degree.

7 Conclusion

In this paper, we introduce a novel task of Multi-modal Sarcasm Generation, aiming to generate sarcastic descriptions like humans. To address the task, we investigate a new dataset, MuSG, containing 5000 images with corresponding sarcastic descriptions. Further, we propose a strong baseline, MTMSG, to benchmark the MuSG dataset. Machine evaluation metrics demonstrate that our proposed MTMSG outperforms various comparison baselines. Moreover, the human evaluation shows that our proposed MSG task is challenging and worth further in-depth research. We consider that MSG opens a new avenue in the domains of sarcasm understanding and generation. In the future, we will explore the detection of key information from the images and the understanding of the intent from the OCR tokens.

Limitations

To better understand the limitations of our proposed MTMSG, we also perform a qualitative error analysis of the incorrectly generated samples. We randomly select 100 incorrectly generated descriptions and find that our model might incorrectly generate those samples mainly due to the misunderstanding of the necessary intent information from the images and OCR tokens. The statistical results reveal that 37% of the incorrectly generated descriptions are caused because the main part of the sarcasm might lie in the images (eg. Figure 5 (a)), while the other 63% error cases are attributed to the failure of our model in capturing the intent information directly from the OCR tokens (eg. Figure 5 (b)). Specifically, in Figure 5 (a), if we want to generate better descriptions, we need to capture the fine-grained visual attribute feature *happy* from the image; In Figure 5 (b), we need to understand the intent information from the OCR tokens that ban guns can make us feel safe when we have lunch in the restaurant. Therefore, to address the above issues in the future, we will further explore the fine-grained key information in the images to help guide the MSG. Besides, we will explore a language interpreter to further understand the key information contained in the OCR tokens.

Acknowledgements

We thank anonymous reviewers for their valuable comments and thoughtful suggestions. This work was supported by the National Natural Science Foundation of China (62276072, 62076100,and 62261003), the Guangxi Natural Science Foundation (No. 2022GXNSFAA035627), Guangxi Scientific and Technological Bases and Talents Special Projects (Application No. 2022AC21300, 2022AC21254), the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, and the Open Research Fund of Key Laboratory of Big Data and Intelligent Robot (SCUT), Ministry of Education.

References

Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. 2014. Word spotting and recognition with embedded attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(12):2552–2566.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Christian Burgers, Margot Van Mulken, and Peter Jan Schellens. 2012. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology*, 31(3):290–310.

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

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4619–4629.

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 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7976–7986.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. 2022. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 367–377. International Committee on Computational Linguistics.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014*, pages 376–380. The Association for Computer Linguistics.

Poorav Desai, Tanmoy Chakraborty, and Md. Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 10563–10571. AAAI Press.

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, NAACL-HLT 2019*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *9th International Conference on Learning Representations, ICLR 2021*.

Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mark James Carman. 2015. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD*.

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 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 5956–5968. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. 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 *ACM MM '21: ACM Multimedia Conference*, 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, ACL 2022*, pages 1767–1777.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization of ACL 2004*, pages 74–81.

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 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 1834–1845.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 6143–6153.

Silviu Oprea, Steven R. Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021*, pages 339–349.

Silviu Vlad Oprea, Steven R. Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 7686–7700. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics, ACL 2002*, pages 311–318.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2249–2255.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 1690–1700.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, NeurIps 2015*, pages 91–99.

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

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12347, pages 742–758.

Jiajia Tang, Kang Li, Ming Hou, Xuanyu Jin, Wanzeng Kong, Yu Ding, and Qibin Zhao. 2022. MMT: multi-way multi-modal transformer for multimodal learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 3458–3465. ijcai.org.

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 2017, NeurIps 2017*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 4566–4575. IEEE Computer Society.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. 2022. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 8164–8175. Association for Computational Linguistics.

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, ACL 2020*, pages 3777–3786.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5076–5088. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the Section "Error Analysis", we describe the limitation of our work.

☒ A2. Did you discuss any potential risks of your work?
No, our work does not involve ethical issues.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, from the section of the Abstract and Introduction, reviewers can easily get the main idea of our work.

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

B ☑ Did you use or create scientific artifacts?

Yes, from the section of Reference, Related Work, "Dataset and Metrics", and Abstract.

☑ B1. Did you cite the creators of artifacts you used?
Yes, from the section of Reference.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No, we only utilized artifacts that are publically avaliable.

☑ 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)?
Yes, from the section of Related Work.

☑ 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?
Yes, from the section of "Dataset and Metrics".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, from the section of Abstract and "Dataset and Metrics".

☑ 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.
Yes, from the section of"Dataset and Metrics".

C ☑ Did you run computational experiments?

Yes, from the section on Experiments and Experimental Results.

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

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?
Yes, from the section of Experiments.

☑ 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?
Yes, from the section of Experiments and Experimental Results.

☑ 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.)?
Yes, from the section of "Dataset and Metrics" and Experiments.

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

Yes, from the section of Experimental Results "Human Evaluation" part.

☑ 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.?
Yes, from the section of Experimental Results "Human Evaluation" part.

☑ 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)?
Yes, from the section of Experimental Results "Human Evaluation" part.

☑ 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?
Yes, from the section of Experimental Results "Human Evaluation" part.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No, there is no ethics review board being involved.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No, we only ask some evaluators for human evaluation.