SimRe: A Simulation of Memes Recreation for Memes Category Detection

Lin Li[1]✉, Leqi Zhong[1], Jian Cui[2], Shaopeng Tang[1], and Xiaohui Tao[3]

[1] Wuhan University of Technology, Wuhan, China
{cathylilin,zlq_lucky,karitown}@whut.edu.cn
[2] Artificial Intelligence Research Institute, iFLYTEK Co., Ltd., Hefei, China
jiancui3@iflytek.com
[3] University of Southern Queensland, Springfield, Australia
Xiaohui.Tao@unisq.edu.au

Abstract. The rapid spread of social media in recent years has given rise to a new multimodal entity: internet memes. The advanced models reconstruct the memes category detection task into an ITM (Image-Text Matching) task based on contrastive learning, showing great improvement over traditional classifiers. However, current methods do not effectively consider memes' characteristics when generating hard negative samples. To address this issue, we propose a memes category detection model via a **Sim**ulation of memes **Re**creation (abbreviated as **SimRe**). Considering the memes distinguished characteristics, i.e., the ease of creation and modification of memes, we design image-text linear interpolation to simulate the recreated process of memes. Then, our be/not prompt templates add the category words of memes into text modality. Based on interpolation and prompt, harder negative samples with a secondary creation style can be generated, bringing perturbation to category label distribution of the original dataset. Finally, experimental results on two datasets show that our model outperforms state-of-the-art methods in terms of accuracy based measures and meanwhile its robustness is enhanced via weight adjustment.

Keywords: Memes Recreation · Linear Interpolation · Negative Samples.

1 Introduction

Internet memes refer to cultural carriers that spread rapidly online, often appearing in a combination of images and text to express the creator's viewpoint and sentiment on specific subject or individuals. For example, Figure 1 shows two internet memes. However, with the increasing complexity of the online environment, originally harmless Internet memes have gradually become the primary carrier of a lot of false and harmful information [12]. To further analyze social media content and strengthen online security management, many researchers have delved into the analysis of internet memes data. Specifically, internet memes category detection refers to determining the category to which memes data belongs

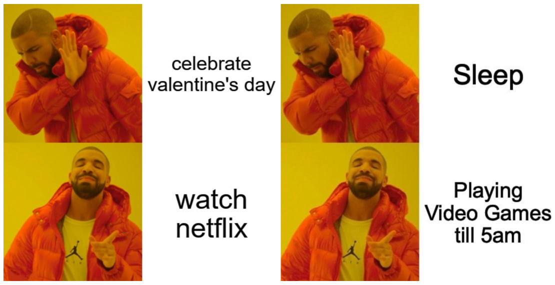


Fig. 1. Internet memes Example from https://www.pinterest.com/sayingimages/pinterests-funniest-memes/

based on its image and text information [16]. The category definition of internet memes is broad, essentially dividing into two types. One is the expression of memes, such as analyzing whether memes data uses a certain propaganda technique [5]. The other focuses on the content of memes, such as harmful or not, analyzing the viewpoints expressed in memes data [13]. Solutions proposed usually treat the memes category detection as a multimodal data classification task [22, 17, 11, 18, 7, 8]. The advanced models reconstruct the memes category detection task into an ITM (Image-Text Matching) task based on contrastive learning [14, 21, 3, 4]. Although the advanced models have achieved great results in accuracy, they still do not pay much attention to the impact of hard negative samples [1]. Specifically, memes often express different content after being spread. In fact, due to the simple and easy-to-modify nature of memes, users will also re-create internet memes based on their own viewpoints and cultural backgrounds during the spread process. This means that existing memes category detection models need to be more applicable to recreated internet memes.

To address the above issues, we propose the **Sim**ulation of memes **Re**creation (**SimRe**) with simulation of memes recreation and image-text contrastive learning. Our propose novel model SimRe is the first attempt simulating the recreated process of memes, which generates more and harder negative samples and brings perturbation to data label distribution. Extensive experiments demonstrate the effectiveness of our model. The results of statistic test also verifies that the superiority of SimRe is statistically significant.

2 Our Model

2.1 Task Definition

Here memes have two modalities: image and text. Usually the memes category detection task is essentially a multimodal classification task. Current contrastive learning based methods show their advantanges under Image-Text Matching framework, such as APCL [4]. We follow them to define our task as below.

Input: text of memes: $[Text_1, \cdots, Text_N]$; image of memes: $[Image_1, \cdots, Image_N]$; label information of memes: $[label_1, \cdots, label_N]$;

Output: predicted memes category: $[\widehat{label}_1, \cdots, \widehat{label}_N]$;

Our task is expressed as a function $f(\cdot)$ that predicts the meme category label through the memes text and image content $<Text, Image>$, as the mapping relationship in Equation 1. The minimization target O of training is in Equation 2. N is total number of samples, and e is usually InfoNCE Loss.

$$< Text, Image > \xrightarrow{f(\cdot)} \widehat{label} \tag{1}$$

$$O = \frac{\sum_{i=1}^{N} e(label_i, \widehat{label}_i)}{N} \tag{2}$$

2.2 Overview

As illustrated in Figure 2, our model has two modules: Simulation of Memes Recreation and Image-text Contrastive Learning.

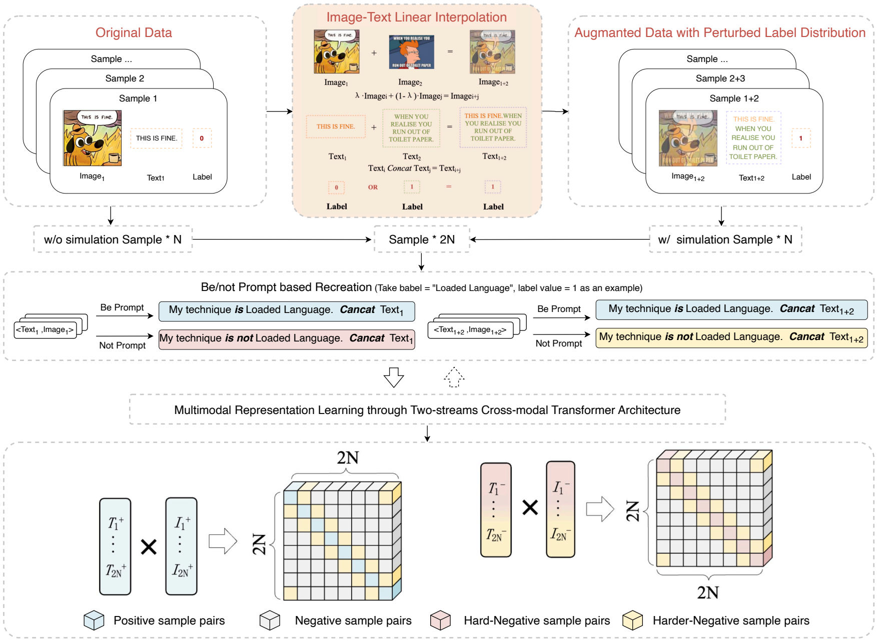

Original Data
Sample ...
Sample 2
Sample 1
THIS IS FINE.
THIS IS FINE.
0
Image$_1$
Text1
Label
Image-Text Linear Interpolation
Image$_1$
Image$_2$
Image$_{1+2}$
λ ·Imagei + (1- λ)·Imagej = Imagei+j
THIS IS FINE.
WHEN YOU REALISE YOU RUN OUT OF TOILET PAPER.
THIS IS FINE.WHEN YOU REALISE YOU RUN OUT OF TOILET PAPER.
Text$_1$
Text$_2$
Text$_{1+2}$
Text$_i$ Concat Text$_j$ = Text$_{i+j}$
0
OR
1
=
1
Label
Label
Label
Augmanted Data with Perturbed Label Distribution
Sample ...
Sample 2+3
Sample 1+2
THIS IS FINE. WHEN YOU REALISE YOU RUN OUT OF TOILET PAPER.
1
Image$_{1+2}$
Text1+2
Label
w/o simulation Sample * N
Sample * 2N
w/ simulation Sample * N
Be/not Prompt based Recreation (Take babel = "Loaded Language", label value = 1 as an example)
<Text$_1$,Image$_1$>
Be Prompt
Not Prompt
My technique is Loaded Language. Cancat Text$_1$
My technique is not Loaded Language. Cancat Text$_1$
<Text$_{1+2}$,Image$_{1+2}$>
Be Prompt
Not Prompt
My technique is Loaded Language. Cancat Text$_{1+2}$
My technique is not Loaded Language. Cancat Text$_{1+2}$
Multimodal Representation Learning through Two-streams Cross-modal Transformer Architecture
T_1^+
T_{2N}^+
I_1^+
I_{2N}^+
2N
2N
T_1^-
T_{2N}^-
I_1^-
I_{2N}^-
2N
2N
Positive sample pairs
Negative sample pairs
Hard-Negative sample pairs
Harder-Negative sample pairs

Fig. 2. The Architecture of Our Proposed SimRe.

The simulation of memes recreation module includes linear interpolation and be/not prompt. First two memes are sequentially extracted from the batch of memes via linear interpolation based recreation. And then their images are linearly fused, texts are concatenated, and label information subjected to an OR operation. Finally, the category words of memes are added into text modality through a prompt template. These operations generate more harder negative samples, causing a perturbation in category label distribution.

The Image-text Contrastive Learning module has multimodal representation learning and training with infoNCE loss. Parameter-shared transformer is adopted to fuse image and text features. The InfoNCE loss is used to optimize image and text representation, thereby increasing the similarity between the image and text of the same sample. Experiments demonstrate that our SimRe can reduce the relative weights of negative samples to maintain its robustness.

2.3 Simulation of Memes Recreation

Linear Interpolation based Recreation: Consider that when memes are recreated on the Internet, the new text is usually changed over the original image, or a new image is used to represent the original text. We simulate this process, combining the image and text information from two memes data to generate new image-text pairs. This operation is based on MixGen [9] and label is generated through an OR operation and combined with newly generated image-text pair to form new memes data. For two memes data $<Text_i, Image_i, Label_i>$ and $<Text_j, Image_j, Label_j>$, $i, j \in \{1, \cdots, N\}$, the equations are defined as follows:

$$Image_{i+j} = \lambda \cdot Image_i + (1 - \lambda) \cdot Image_j \tag{3}$$

$$Text_{i+j} = Text_i \ Concat \ Text_j \tag{4}$$

$$Label_{ij} = Label_i \ OR \ Label_j \tag{5}$$

λ is a hyperparameter between 0 and 1. $Concat$ means concatenating two text information. OR operation merges two labels. From Figure 2, the upper example shows that most of the image information is kept and the text words are completely remained. From the semantic meaning, the generated sample may be added incrementally something new since two text sentences work together. Meanwhile, these new image-text pairs are with semantic relationships preserved, which ensures the correctness of the label information.

With these operations, the label distribution is perturbed, which has a positive impact on the robustness of our model. Briefly speaking, the image-text feature similarity between the newly generated memes and the original memes will be significantly higher than that of other completely mismatched image-text pairs. Therefore, more and harder negative samples can be generated.

Be/Not prompt based Recreation: As shown in the second part of Figure 2, following Cui et al. [4], image-text matching modelling incorporates category words into text modality in a similar writing style to memes themselves. $<Text, Image>$ represents the image-text pairs belong to a memes category as detection target, and $<\overline{T}ext, \overline{I}mage>$ represents the image-text pairs not belong to the targeted memes category.

All the generated data and original data are together prompted in this module. We design different prompts for each task and mark new text with **is/is not** as $Text_i^{be/not}$, so the pairs prompt as follows:

$$< Text, Image > \xrightarrow{be/not\ prompt} \quad < Text^{be/not}, Image > \tag{6}$$

$$< \overline{T}ext, \overline{I}mage > \xrightarrow{be/not\ prompt} \quad < \overline{T}ext^{be/not}, \overline{I}mage > \tag{7}$$

After these image-text pairs are obtained, they are processed by the modality encoder such as BLIP [10], and the embeddings are fed to our Image-Text Contrastive Learning module.

2.4 Image-Text Contrastive Learning

Multimodal Representation Learning. We adopt a two-streams cross-modal Transformer architecture to learn joint embeddings. Other multimodal pre-trained models, such as CLIP [15] or BLIP [10], can also be used as encoders. The image-text representation pairs obtained in the previous step will be separately fed into the corresponding two-streams cross-modal Transformer as *Sources* and *Targets*. Specifically, Equation 8 shows the details in the third part of Figure 2.

$$\begin{cases} I^{be/not} = TF(sources : e(Text^{be/not}), targets : e(Image)) \\ T^{be/not} = TF(sources : e(Image), targets : e(Text^{be/not})) \\ \overline{I}^{be/not} = TF(sources : e(\overline{T}ext^{be/not}), targets : e(\overline{I}mage)) \\ \overline{T}^{be/not} = TF(sources : e(\overline{I}mage), targets : e(\overline{T}ext^{be/not})) \end{cases} \tag{8}$$

$TF()$ represents $Transformers$, $e()$ represents the embedding results. Finally, this process will generate four image-text features pairs, as $<T^{be}_{1\cdots K}, I^{be}_{1\cdots K}>$, $<\overline{T}^{not}_{1\cdots Q}, \overline{I}^{not}_{1\cdots Q}>$, $<T^{not}_{1\cdots K}, I^{not}_{1\cdots K}>$, $<\overline{T}^{be}_{1\cdots Q}, \overline{I}^{be}_{1\cdots Q}>$. Then these features pairs will be send to the next step.

Training with InfoNCE Loss. In this process, the generated image-text features pairs will be recomposed into two types, marked as $<T^{+}_{1\cdots 2N}, I^{+}_{1\cdots 2N}>$ and $<T^{-}_{1\cdots 2N}, I^{-}_{1\cdots 2N}>$. The combinations are defined as follows:

$$< T^{be}_{1\cdots K}, I^{be}_{1\cdots K} >, < \overline{T}^{not}_{1\cdots Q}, \overline{I}^{not}_{1\cdots Q} > \quad \Longrightarrow < T^{+}_{1\cdots 2N}, I^{+}_{1\cdots 2N} > \tag{9}$$

$$< T^{not}_{1\cdots K}, I^{not}_{1\cdots K} >, < \overline{T}^{be}_{1\cdots Q}, \overline{I}^{be}_{1\cdots Q} > \quad \Longrightarrow < T^{-}_{1\cdots 2N}, I^{-}_{1\cdots 2N} > \tag{10}$$

Among them, $<T^{+}_{1\cdots 2N}, I^{+}_{1\cdots 2N}>$ represents the content of memes and prompt text are consistent, while the other represents inconsistent. The above two image-text pairs will be split separately when training with InfoNCE Loss. The model will calculate the similarity between each image feature and text feature through the dot product method. Specifically, for the image-text feature pair $<T^{+}_{1\cdots 2N}, I^{+}_{1\cdots 2N}>$, SimRe will calculate the similarity between each $T^{+}_{1\cdots 2N}$ and all $I^{+}_{1\cdots 2N}$ to obtain a similarity with a dimension of $2N * 2N$ matrix. Then, we can get a similarity matrix with a dimension of $4N * 2N$ by splicing them together.

Three different samples are defined as follows:

Positive samples: Image and text come from the same memes sample, and the prompt template content in the text is the same as the memes label.

Hard negative samples: Image and text come from the same memes but the prompt template content in the text is different from the memes label.

Negative samples: Image and text do not come from the same memes.

InfoNCE loss is commonly used in contrastive learning. It can effectively capture semantic information via maximizing the similarity of positive samples and minimizing the similarity of negative samples. Detailed information is defined as Equation 11 and Equation 12.

$$Loss_i = -log\frac{exp(T_i^+ \cdot I_i^+/\tau)}{exp(T_i^+ \cdot I_i^+/\tau) + U_i} \tag{11}$$

$$U_i = \sum_{j\in[\![1,N]\!],j\neq i} exp(T_i^+ \cdot I_j^+/\tau) \quad + \sum_{k\in[\![1,N]\!]} exp(T_i^- \cdot I_k^-/\tau) \tag{12}$$

τ represents the temperature coefficient in contrastive learning that is used to affect the model's discrimination of negative samples. Adjusting this coefficient helps to balance sensitivity of samples in contrastive learning.

3 Experiments

Our study aims to answer two research questions: Can our SimRe achieve competitive performance facing label distribution perturbation? How do hard negatives affect detection performance?

3.1 Dataset and Implementation Details

Both of propaganda detection [5–7] and harmfulness memes detection [13, 14] are adopted to verify the effectiveness of our SimRe. We use SemEval-2021 Task 6 dateset [6] for propaganda detection and Harm-P dataset [14] for harmfulness memes detection, which is both publicly available. Evaluation metrics include widely used F1 and Accuracy. The SimRe uses BLIP as the image and text encoder of memes, specifically uses ViT for image encode and BERT for text encode. The image fusion hyperparameter λ is set to 0.5. The temperature coefficient τ of the InfoNCE loss function is set to 0.2 for memes propaganda technique category detection and 0.5 for harmfulness category detection. Adam is used as the optimizer, and its learning rate is set to 5e-5. Steplr is used as the scheduler to adjust the learning rate of the optimizer.

3.2 Overall Performance

It can be seen in Sec. 2.3 that our simulation operations generate more harder negative samples, and cause a perturbation in the original dataset's category label distribution. In the case of inconsistent distribution of training and test dataset, it becomes more difficult to train robust models. For evaluation, our SimRe is compared with various memes category detection models and advanced contrastive learning models on two datasets, as shown in Table 1.

Table 1. Overall Experimental Results

Model	Venue	Propaganda technique detection			Harmfulness memes detection		
		F1-Micro	F1-Macro	p-value	ACC	F1	p-value
ERNIE-VIL [20]	AAAI 2021	58.1	27.31	0.0036	78.54	59.65	0.0315
CLIP [15]	ICML 2021	55.81	24.04	0.0109	77.00	56.85	0.0444
BLIP [10]	ICML 2022	57.34	28.32	0.0031	79.00	59.84	0.0302
MOMENTA [14]	EMNLP 2021	58.24	26.44	0.0046	87.14	66.66	0.0043
TOT [21]	AAAI 2023	-	-	-	88.61	71.54	0.0022
APCL [4]	ICASSP 2023	58.72	30.76	-	-	-	-
MViTO-GAT [2]	MTA 2024	54.18	20.99	-	-	-	-
SimRe-CLIP (Ours)		59.87	31.88	0.0017	89.86	71.64	0.0008
SimRe-BLIP (Ours)		**59.98**	**32.00**	-	**89.96**	**71.68**	-
w/o Simulation		59.58	31.71	0.0019	88.81	70.72	0.0020

SimRe shows competitive performance with state-of-the-art in terms of F1 and Accuracy although simulation brings perturbation to label distribution of original dataset. Our simulation operations cause inconsistent distribution of training and test dataset, which may lead to decreased performance. However, it can be seen in Table 1 that SimRe shows competitive performance in both propaganda technique detection task and harmfulness memes detection task. **In addition, SimRe has significant statistical differences compared to other models.** In order to further verify whether the difference between our model and other models belongs to random fluctuations, we use the p-value in the T test for testing. It can be seen from Table 1 that all p-value between other models and our model are less than 0.05, indicating that the superiority of SimRe is statistically significant.

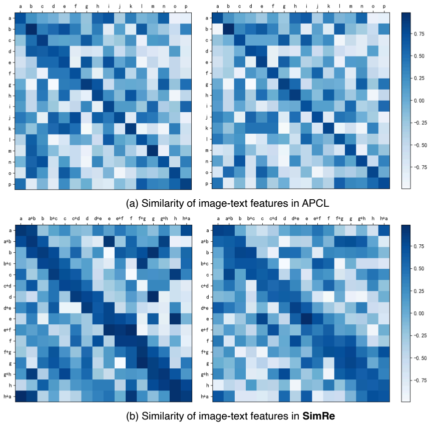
Overall, by introducing simulation of memes recreation and bringing perturbation to the label distribution of data samples, SimRe can learn more subtle differences in meme content and be more effective.

3.3 The Effect Analysis of Hard Negatives

According to the advanced theoretical research [19], the robustness of our SimRe is enhanced through generating abundant hard negative samples. The following experimental results demonstrate that our SimRe can reduce the relative weight of negative samples with the highest similarity to further enhance robustness.

Similarity of samples. We have visualized the similarity of the image and text features in the contrastive learning module of the two advanced models. As shown in Figure 3, the left side is the similarity of features between T^+ and I^+, while the right side is the similarity between T^- and I^-. The darker color means more similar. It can be seen that the number of hard negative samples with greater similarity in our model has greatly increased. Specifically, the number of hard negative samples in our model is 4 times more than APCL.

Robust analysis with relative weights. The weights of negative samples in contrastive learning is assigned according to $exp[f_\theta/\tau]$, which indicates that the negative sample with the highest similarity will receive the highest weight [19]. Therefore, reducing the weights of negative samples is a key optimal



(a) Similarity of image-text features in APCL

(b) Similarity of image-text features in **SimRe**

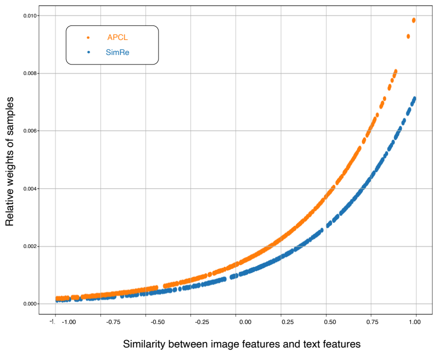


Fig. 3. The number of negatives in SimRe is 4 times more than APCL

Fig. 4. Relative Weights of Negatives

direction to further improve the robustness of the model. Our model reduces the relative weight through generating more and harder negative samples and bringing perturbation to label distribution. In Figure 4, we normalized the weights of samples within a certain batch and visualized the relative weights to see the changes more intuitively. The difference in relative weights between SimRe and APCL is more obvious in terms of greater similarity, which means SimRe pays less attention to samples with higher similarity, and enhances the robustness.

Temperature coefficient τ. In addition, in the InfoNCE loss function, the temperature coefficient is a key hyperparameter, which is used to control the scale of the similarity score, thus affecting the size of the gradient and the ease of optimization. A lower temperature coefficient will increase the scale of the model output, making the loss function more sensitive to positive sample pairs and making the distinction between negative samples more obvious. The temperature coefficient in contrastive learning is usually set between 0.05 and 0.5. We conducted experiments on four different temperature coefficient to explore the model performance. The results are shown in Table 2.

Table 2. Temperature Coefficient Analysis

τ	**Propaganda technique detection**		**Harmfulness memes detection**	
	F1-Micro	F1-Macro	F1	ACC
0.05	59.86	31.99	89.88	71.55
0.1	59.90	31.97	89.94	71.58
0.2	**59.91**	**32.00**	89.92	71.60
0.5	59.88	31.94	**89.96**	**71.68**

The temperature coefficient is not well positively correlated with F1 scores. The third row and the fourth row show that our SimRe takes the best setting 0.2 for propaganda technique detection and 0.5 for harmfulness memes detection. Reducing the temperature coefficient can increase the gap between samples, thereby improving the model's ability to learn the difference between features. However, when the temperature coefficient is small, the relative weight of each sample in contrastive learning will be increased. Overall, an appropriate temperature coefficient for training is important.

4 Conclusions

This paper proposes a novel memes category detection model named Simulation of memes Recreation(SimRe), which can bring perturbation to category label distribution of the original dataset and adjust the relative weights in contrastive learning by generating hard negative samples, thereby further enhancing performance. A series of analytical and empirical comparisons are conducted to show its effectiveness in terms of accuracy and robustness.

Acknowledgements

This work is supported by NSFC, China (No. 62276196), Guangxi Science and Technology Major Program (No.AA24206067).

References

1. Robinson, J.D., Chuang, C., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
2. Chen, P., Zhao, L., Piao, Y., Ding, H., Cui, X.: Multimodal visual-textual object graph attention network for propaganda detection in memes. Multim. Tools Appl. **83**(12), 36629–36644 (2024)
3. Cui, J., Li, L., Tao, X.: Be-or-not prompt enhanced hard negatives generating for memes category detection. In ICME 2023. pp. 174–179 (2023)
4. Cui, J., Li, L., Zhang, X., Yuan, J.: Multimodal propaganda detection via anti-persuasion prompt enhanced contrastive learning. In ICASSP 2023. pp. 1–5 (2023)
5. Dimitrov, D., Ali, B.B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Martino, G.D.S.: Detecting propaganda techniques in memes. In ACL/IJCNLP 2021. pp. 6603–6617 (2021)
6. Dimitrov, D., Ali, B.B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., Martino, G.D.S.: Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In SemEval@ACL/IJCNLP 2021. pp. 70–98 (2021)
7. Feng, Z., Tang, J., Liu, J., Yin, W., Feng, S., Sun, Y., Chen, L.: Alpha at semeval-2021 task 6: Transformer based propaganda classification. In SemEval@ACL/IJCNLP 2021. pp. 99–104 (2021)
8. Ghadery, E., Sileo, D., Moens, M.: LIIR at semeval-2021 task 6: Detection of persuasion techniques in texts and images using CLIP features. In SemEval@ACL/IJCNLP 2021. pp. 1015–1019 (2021)

9. Hao, X., Zhu, Y., Appalaraju, S., Zhang, A., Zhang, W., Li, B., Li, M.: Mixgen: A new multi-modal data augmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023. pp. 379–389 (2023)
10. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pretraining for unified vision-language understanding and generation. In ICML 2022. pp. 12888–12900 (2022)
11. Li, P., Li, X., Sun, X.: 1213li at semeval-2021 task 6: Detection of propaganda with multi-modal attention and pre-trained models. In SemEval@ACL/IJCNLP 2021. pp. 1032–1036 (2021)
12. Martino, G.D.S., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R.D., Nakov, P.: A survey on computational propaganda detection. In IJCAI 2020. pp. 4826–4832 (2020)
13. Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M.S., Nakov, P., Chakraborty, T.: Detecting harmful memes and their targets. In ACL/IJCNLP 2021. pp. 2783–2796 (2021)
14. Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M.S., Nakov, P., Chakraborty, T.: MOMENTA: A multimodal framework for detecting harmful memes and their targets. In EMNLP 2021. pp. 4439–4455 (2021)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In ICML 2021. pp. 8748–8763 (2021)
16. Sharma, S., Alam, F., Akhtar, M.S., Dimitrov, D., Martino, G.D.S., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., Chakraborty, T.: Detecting and understanding harmful memes: A survey. In IJCAI 2022. pp. 5597–5606 (2022)
17. Singh, P., Lefever, E.: LT3 at semeval-2021 task 6: Using multi-modal compact bilinear pooling to combine visual and textual understanding in memes. In SemEval@ACL/IJCNLP 2021. pp. 1051–1055 (2021)
18. Tian, J., Gui, M., Li, C., Yan, M., Xiao, W.: Mind at semeval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In SemEval@ACL/IJCNLP 2021. pp. 1082–1087 (2021)
19. Wu, J., Chen, J., Wu, J., Shi, W., Wang, X., He, X.: Understanding contrastive learning via distributionally robust optimization. In NeurIPS 2023.
20. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In AAAI 2021. pp. 3208–3216 (2021)
21. Zhang, L., Jin, L., Sun, X., Xu, G., Zhang, Z., Li, X., Liu, N., Liu, Q., Yan, S.: Tot : topology-aware optimal transport for multimodal hate detection. In AAAI 2023. pp. 4884–4892 (2023)
22. Zhu, X., Wang, J., Zhang, X.: YNU-HPCC at semeval-2021 task 6: Combining ALBERT and text-cnn for persuasion detection in texts and images. In emEval@ACL/IJCNLP 2021. pp. 1045–1050 (2021)