Identifying Creative Harmful Memes via Prompt based Approach

Junhui Ji
juji7690@uni.sydney.edu.au
School of Computer Science,
University of Sydney
Sydney, Australia

Wei Ren
wren2033@uni.sydney.edu.au
School of Computer Science,
University of Sydney
Sydney, Australia

Usman Naseem
usman.naseem@sydney.edu.au
School of Computer Science,
University of Sydney
Sydney, Australia

ABSTRACT

The creative nature of memes has made it possible for harmful content to spread quickly and widely on the internet. Harmful memes can range from spreading hate speech promoting violence, and causing emotional distress to individuals or communities. These memes are often designed to be misleading, manipulative, and controversial, making it challenging to detect and remove them from online platforms. Previous studies focused on how to fuse visual and language modalities to capture contextual information. However, meme analysis still severely suffers from data deficiency, resulting in insufficient learning of fusion modules. Further, using conventional pretrained encoders for text and images exhibits a greater semantic gap in feature spaces and leads to low performance. To address these gaps, this paper reformulates a harmful meme analysis as an auto-filling and presents a prompt-based approach to identify harmful memes. Specifically, we first transform multimodal data to a single (i.e., textual) modality by generating the captions and attributes of the visual data and then prepend the textual data in the prompt-based pre-trained language model. Experimental results on two benchmark harmful memes datasets demonstrate that our method outperformed state-of-the-art methods. We conclude with the transferability and robustness of our approach to identify creative harmful memes.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision representations**.

ACM Reference Format:
Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying Creative Harmful Memes via Prompt based Approach. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543507.3587427

1 INTRODUCTION

Memes have become ubiquitous in internet culture, widely shared and enjoyed as a form of humor and expression. Creative memes, in particular, have emerged as a unique genre characterized by their originality and artistic value. However, while many memes can be harmless and even creative expressions of humor, others can be harmful and spread toxic messages and ideologies [1]. Harmful memes can cause real-world harm, spreading false or dangerous information or normalizing harmful attitudes [18, 23, 25].

Previous research to identify harmful memes has primarily used traditional pretrained encoders to derive the image and text representations and focused on designing new methods to fuse multimodal data [11] and minimize the gap between modalities to capture the semantic and contextual information (Figure 1(a)). However, leveraging these traditional pretrained encoders results in divergent image and text representation spaces and makes it challenging to model any relationship between them, leading to a semantic gap in understanding the meaning of a meme [28].

Different modalities can be combined at various levels. Early fusion approaches [8, 14, 16] combine raw inputs, such as image and text, to learn a joint representation of both modalities. Late fusion approaches [9], on the other hand, learn end-to-end models for each modality and concatenate their outputs before making predictions. A recent multimodal framework for harmful memes called MOMENTA, presented by Pramanick et al. [20], uses cross-modality attention fusion by concatenating text and image features and learning a cross-modal weight matrix to modulate the concatenated features. While these fusion-based models have improved performance, they may not be suitable for hateful meme analysis because the text in a meme does not necessarily function as an image caption, and text and image may imply different meanings, which challenges the assumption that the associated text describes the contents of the image [11].

In addition, collecting and annotating considerable harmful memes is challenging, particularly for the fine-grained classes [30]. Considering the cost, variability, and other relevant time or hardware constraints for labeling, identifying harmful memes still suffer from data limitation, leading to inadequate learning of fusion modules and poor performance [11].

To address the above limitations and motivated by the success of prompt learning in natural language processing (NLP) [3, 12, 15], we attempt to introduce a prompt-based learning-based method to identify harmful memes. Our main contributions in this paper are as follows:

- We reformulate harmful meme analysis as an auto-filling task. To our best knowledge, this is the first attempt to identify harmful memes using the prompt paradigm.
- We present a novel approach that transforms multimodal data into a single modality to address the semantic gap in fusion-based methods before feeding it to prompt-based learning to address the data scarcity issue for a harmful meme analysis.

First two authors contributed equally.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3587427

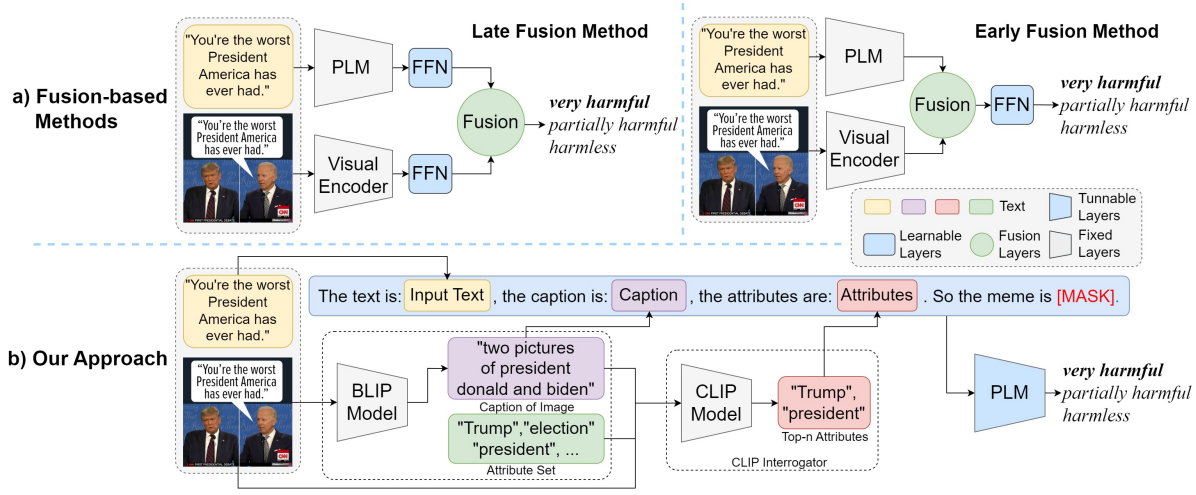

Figure 1: The architectures of (a) fusion based approaches and (b) our approach.

- We show that our approach outperforms state-of-the-art baselines and discuss the transferability and robustness of our approach.

2 PROPOSED APPROACH

Overview: Our approach includes the following modules: (i) an image-to-text module where we first transform an image to text to capture better semantics of a meme; and (ii) a prompt-based learning module to address the data scarcity. Figure 1 (b) demonstrates the overall architecture of our proposed approach. In the subsequent discussion, we explain each module in detail.

Image-to-text: We first transform an image into a text to reduce a semantic gap between an image (I_i) and text (T_i) of a given meme (M). To do so, we generated captions of visual content within a meme using BLIP [13], a vision-language pre-training framework that transfers flexibly to vision-language understanding and generation tasks. We uniformly resized an image before feeding it to the BLIP to generate a caption (I_i^C) (equation 1).

$$I_i^C = BLIP(I_i) \tag{1}$$

In addition to captions, following previous studies [4, 20, 26], to enhance the ability of the proposed approach to detect harmful content and contextualize the input meme, attributes are introduced as high-level image concepts. Figure 1 (pink box) demonstrates that the identified image attributes, such as 'trump' and 'president' capture the proper context of a meme.

To find out the top N attributes, we used a CLIP Interrogator[1] that combines CLIP [22], a pre-trained visual-linguistic model and BLIP [13] to optimize text prompt to match a given image. Using CLIP Interrogator, we computed the similarity score using the cosine similarity by comparing an image I_i and each attribute $k_j \in K$ (equation 2).

$$I_i^K = N_{k_j \in K}\{cos(CLIP[I_i, k_j])\} \tag{2}$$

where N is the number of top attributes, and K is the set of common entities and attribute set.

As an output of our image-to-text module, multimodal data (i.e., an image and text within a meme) is transformed into a single modality (i.e., text) that consists of an input text (T_i), caption (I_i^C) and attributes (I_i^K). This transformation step addresses the semantic gap in the feature space and results in a better understanding of a harmful meme.

Prompt-based-learning: To address the data scarcity issue, we leveraged prompt based learning, where we fine-tuned a pretrained language model (PLM). The downstream tasks, given a PLM (M), are converted into a masked language modeling (MLM) task with new label space and various prompt templates.

Specifically, we need to define a prompt template $\mathbf{T}(T_i, I_i^C, I_i^K)$ and a label mapping or verbalizer. $\mathbf{T}$ incorporates supplementary prompt tokens as contextual cues for the task. Typically, a `[MASK]` token is retained in $\mathbf{T}$. We provide $\hat{\mathbf{T}}$ = [CLS] $\mathbf{T}$ [SEP] as input to M, allowing the pre-trained MLM head to predict a potential token at the `[MASK]` position. More specifically, we constructed a template $\mathbf{T}$ for our task as:

$$\mathbf{T}(T_i, I_i^C, I_i^K) = \text{The text is: } T_i, \text{ the caption is: } I_i^C, \text{ the attributes are: } I_i^K. \text{ So the meme is [MASK].}$$

where T_i represents the input text, I_i^C as captions, and I_i^K as attributes, then let M decide what is most suitable to fill in `[MASK]`. We optimize our approach by minimizing cross-entropy loss.

3 EXPERIMENTS

Experimental settings: We trained our approach and all baselines using Pytorch on NVIDIA Tesla T4 GPU, with 16GB dedicated GPU memory, with CUDA-11.2 and cuDNN-8.1.1 installed. For the BLIP model, we used official settings of BLIP w/ ViT-L

[1] https://github.com/pharmapsychotic/clip-interrogator

for the image captioning task[2]. For the CLIP model, we use official settings of CLIP ViT-L/14[3]. Our code is publicly available at https://github.com/jasonnoy/H-meme-www.

Parameter settings We used T5-base[4] as our backbone model. For baseline models Text-Prompt, PVLM and our proposed model, we used the same parameters of batch size = 30, epoch number = 30 and learning rate = 0.0001. We selected the top 4 attributes in our experimental settings and used OpenPrompt[5] for prompt learning. We used Adam [10] as the optimizer and cross-entropy [17] as the loss function.

Datasets: We used Harm-C [20] and Harm-P [20], which consist of harmful memes related to COVID-19 and US politics (Table 1).

Table 1: Statistics of the datasets used

Dataset	Split	#Memes	#Very Hamrful	#Partially Harmful	#Harmless
Harm-C	Train	3,013	182	882	1,949
	Validation	177	10	51	116
	Test	354	21	103	230
	Total	3,544	213	1,036	2,295
Harm-P	Train	3,020	216	1,270	1,534
	Validation	177	17	69	91
	Test	355	25	148	182
	Total	3,552	258	1,487	1,807

3.1 Evaluation metrics

We used the evaluation metrics widely used in previous similar studies [20]. Specifically, we used Accuracy, Macro-F1, and Macro-Averaged Mean Absolute Error (MMAE) [2] to evaluate the effectiveness of the proposed approach.

Baselines: We compared our method with unimodal (text and image only) and multimodal-based state-of-the-art (SOTA) methods. For text only, we used TextBERT [5] and a prompt-based method (i.e., BERT+prompt) [19]. For image only, we used VGG19 [24], DensNet-161 [7], ResNet-152 [6], and ResNeXt-101 [27]. For multimodal, we used different fusion strategies to fuse visual and textual features. Specifically, we used Late Fusion, Concat-BERT, MMBT [8], ViLBERT [16], VisualBERT [14], CLIP [21], a prompt-based method i.e., PVLM [29] and MOMENTA [20].

4 RESULTS AND ANALYSIS

Comparison with Baselines: Table 2 shows the overall results of our proposed approach when evaluated against state-of-the-art (SOTA) methods on the two datasets. As expected, using text-only baseline methods yields better performance than image-only baseline methods since text contains more explicit information than images. Nonetheless, we observe that the performance of both the text-only and image-only models is unsatisfactory, as neither exceeds an F1-score of 53.59% on Harm-C or 55.09% on Harm-P.

For multimodal baseline methods, fusion-based methods demonstrate superior performance compared to models that only utilize text or visuals alone (as shown in Table 2). For example, the most basic form of fusion, known as late fusion, slightly outperforms unimodal models. However, the SOTA multimodal baseline methods such as ViLBERT, VisualBERT, MMBT, and CLIP, which were specifically designed for vision-and-language tasks, perform inadequately when compared to MOMENTA (the best baselines) which were created to detect harmful memes and achieved an F1-score of 54.74% on the Harm-C dataset and 66.66% on the Harm-P dataset. We attribute this low performance of fusion-based multimodal methods to the data scarcity issue that leads to insufficient learning.

Table 2: Results: Proposed approach v/s the baselines.

Modality	Model	Harm-C			Harm-P		
		Acc	F1	MMAE	Acc	F1	MMAE
Text only	BERT	68.93	48.72	0.5591	74.55	54.08	0.7742
	BERT+Prompt	71.88	53.59	0.6837	55.37	55.09	0.8976
Image only	VGG19	66.24	41.76	0.6487	73.65	51.89	0.8466
	DenseNet-161	65.21	42.15	0.6326	71.80	50.98	0.8388
	ResNet-152	65.29	43.02	0.6264	71.02	50.64	0.8900
	ResNeXt-101	66.55	43.68	0.6499	71.84	51.45	0.8422
Multimodal	Late-Fusion	66.67	45.06	0.6077	76.20	55.84	0.7245
	Concat-BERT	65.54	43.37	0.5976	76.04	55.95	0.7450
	MMBT	68.08	50.88	0.6474	78.14	58.03	0.7008
	ViLBERT	75.71	48.82	0.5329	84.66	64.70	0.6982
	VisualBERT	74.01	53.85	0.5303	84.02	63.68	0.7020
	CLIP	67.04	44.25	0.6228	77.00	56.85	0.7852
	PVLM	59.47	38.96	0.8532	67.01	65.72	0.7313
	MOMENTA	77.10	54.74	0.5132	87.14	66.66	0.6805
	Proposed	80.23	60.00	0.5762	88.73	89.64	0.1970

We observe that the proposed approach outperformed all baselines with an F1-Score of 60% (an absolute increase of 5.26%) on Harm-C and 89.64% (an absolute increase of 22.98%) on Harm-P compared to MOMENTA, which is designed to identify harmful memes. We also note that other multimodal approaches (i.e., MMBT, ViLBERT, VisualBERT, and CLIP) designed for other tasks are less desirable in detecting harmful memes.

Ablation analysis: As demonstrated in Table 3, each module in the proposed approach contributed to the overall performance of both datasets in the ablation analysis. For example, when we removed captions, attributes, or both (i.e., no captions and attributes) from our proposed approach, there was a remarkable degradation in F1-score on both tasks. Similarly, an F1-score dropped when we removed text. Hence, we conclude that combining all modules adds to improved performance.

Table 3: Ablation analysis

Text	Caption	Attributes	Harm-C	Harm-P
✓	✓	✓	60.00	89.64
✓	✓	x	58.07	88.87
✓	x	✓	54.59	83.52
✓	x	x	53.55	70.97
x	✓	✓	52.42	68.99

Transferability: Similar to previous study [20], we also performed the transferability test to evaluate the effectiveness and robustness of our approach (Table 4). It is clear from the results that our approach consistently outperformed the best baselines (MOMENTA)

[2]https://github.com/salesforce/BLIP, model checkpoint
[3]https://huggingface.co/openai/clip-vit-large-patch14
[4]https://huggingface.co/t5-base
[5]https://github.com/thunlp/OpenPrompt

and achieved the best F1-score. This transferability demonstrates the robustness and generalizability of the our approach.

Table 4: Transferability test of the proposed model compared with MOMENTA on dataset Harm-Covid, on Harm-Policitics, and the combination.

Dataset	Model	Harm-C	Harm-P	Combined
Harm-C	MOMENTA	54.74	41.91	45.23
	Proposed	60.00	45.55	68.54
Harm-P	MOMENTA	31.47	66.66	46.29
	Proposed	45.05	89.64	64.45
Combined	MOMENTA	40.11	54.93	47.47
	Proposed	52.45	89.60	68.44

Effect of a different N: We also investigated the effects of the number of attributes (N) by increasing the value from 1 to 20 (Figure 2). We observed that with the increase of the N value, the model performance increases first and then decreases. When *N* is large, it leads the low performance because of injecting the noise. Our findings show that the value of (N) = 4 delivers the highest model performance on both datasets.

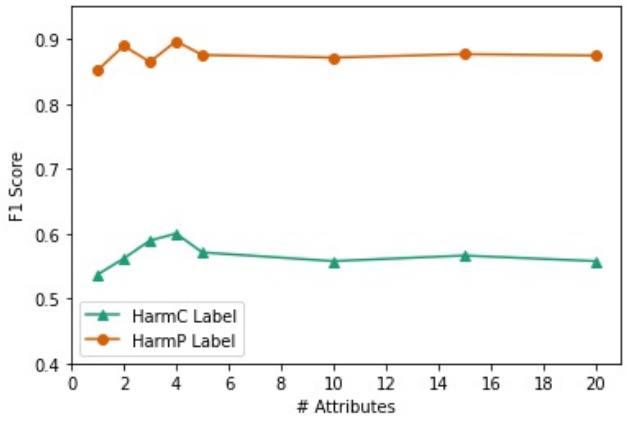

Figure 2: Effect of different N values

Qualitative analysis: The images in Figure 3 are instances of memes that were accurately predicted by our approach. We attribute this success to the ability of our approach to reduce the semantic divide between the image and the accompanying text, allowing our model to grasp the contextual details and improve its overall effectiveness. From these examples, it is evident that our approach, which transforms multimodal data into a singular modality to bridge the semantic gap prior to utilizing prompt-based learning to capture contextual information, performs better than unimodal and multimodal baselines. Moreover, it even surpasses the current state-of-the-art method for analyzing harmful memes.
Error analysis: We present examples where proposed approach failed: (i) the captions generated provide insufficient information; and (ii) memes that need additional background knowledge to understand and correctly predict (Figure 4). In Figure 4, we showed

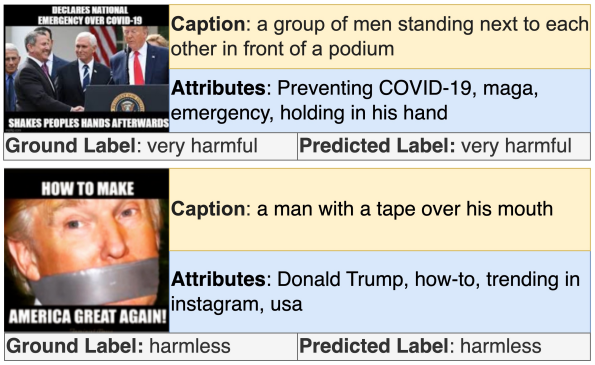

Figure 3: Qualitative analysis: Examples of correctly predicted memes with generated caption and attributes

two examples of incorrectly predicted memes. We postulate this incorrect prediction for the following reasons. First, in the first example (Figure 4 - top), the caption generated did not capture the context of a meme, i.e., the people in the image are looking up into the sun. Further, for attributes, the most significant figure, "Donald Trump", in the image is captured; however, the rest of the attributes are misleading and not informative and lead to the wrong prediction. Similarly, in the second example (Figure 4 - bottom), the wrong prediction is due to the poor quality of generated caption, and the attributes are not informative.

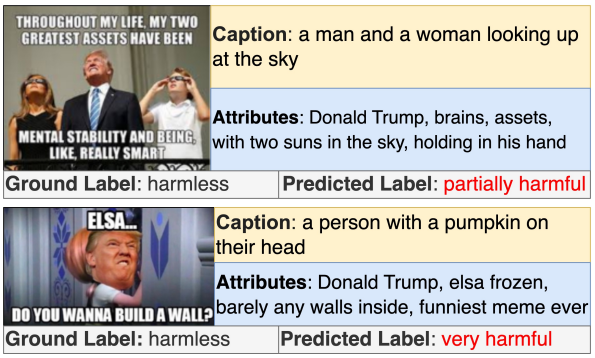

Figure 4: Error analysis: Incorrectly predicted examples with generated caption and attributes

5 CONCLUSION

Creative harmful memes have become a prevalent phenomenon in the digital age and often spread rapidly through social media and can significantly impact individuals and society as a whole. In this work, we reformulated the analysis of creative harmful memes as an auto-filing task to avoid more inputs and reduce a semantic gap to capture contextual information in understanding a meme. We presented a novel prompt based approach, which transforms multimodal data into a single modality to address the semantic gap in fusion-based methods before feeding it to prompt-based learning to address the data scarcity issue and identify harmful memes. Experimental results demonstrated that the proposed approach outperformed the state-of-the-art baseline methods.

REFERENCES

[1] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541* (2021).

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *2009 Ninth international conference on intelligent systems design and applications*. IEEE, 283–287.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

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

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[8] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).

[9] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33 (2020), 2611–2624.

[10] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[11] Gokul Karthik Kumar and Karthik Nanadakumar. 2022. Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. *arXiv preprint arXiv:2210.05916* (2022).

[12] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3045–3059.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://doi.org/10.48550/ARXIV.2201.12086

[14] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021).

[16] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[17] Shie Mannor, Dori Peleg, and Reuven Rubinstein. 2005. The cross entropy method for classification. 561–568. https://doi.org/10.1145/1102351.1102422

[18] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A Multimodal Framework for the Identification of Vaccine Critical Memes on Twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 706–714.

[19] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases? https://doi.org/10.48550/ARXIV.1909.01066

[20] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184* (2021).

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/ARXIV.2103.00020

[23] Shivam Sharma, Firoj Alam, Md Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. 2022. Detecting and Understanding Harmful Memes: A Survey. *arXiv preprint arXiv:2205.04274* (2022).

[24] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*. 1–6.

[26] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems?. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 203–212.

[27] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[28] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. 2022. Prompting for Multi-Modal Tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3492–3500.

[29] Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[30] Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified Multi-modal Pre-training for Few-shot Sentiment Analysis with Prompt-based Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 189–198.