A Multimodal Framework for the Identification of Vaccine Critical Memes on Twitter

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

Jinman Kim
jinman.kim@sydney.edu.au
School of Computer Science, University of Sydney
Sydney, Australia

Matloob Khushi
matloob.khushi@brunel.ac.uk
Department of Computer Science, Brunel University
London, UK

Adam G. Dunn
adam.dunn@sydney.edu.au
School of Medical Sciences, University of Sydney
Sydney, Australia

WSDM2023

Code:None.

2023. 6. 10 • ChongQing

gesis
Leibniz-Institut
für Sozialwissenschaften

Reported by Yang Peng

1.Introduction

2.Method

3.Experiments

Introduction

Figure 1: Examples of vaccine critical memes. Note that in a meme shown on the left, an image becomes a humorous way to identify that a meme is vaccine critical, whereas, for a meme on the right, a text suggests that a meme is vaccine critical.

Problem:

While previous work may not be able to **capture global and local representations** of both textual and visual content within memes and fails to **capture contextual information**。

Contributions:

- We release a manually annotated dataset of 10,244 memes to identify vaccine critical memes on Twitter.
- We present a multimodal framework that learns global and local representations of visual and textual content within memes and captures contextual information.
- We show that the proposed multimodal framework outperforms state-of-the-art baselines with an F1-Score of 84.2% and also establish the transferability and generalis ability of the proposed framework.

Method

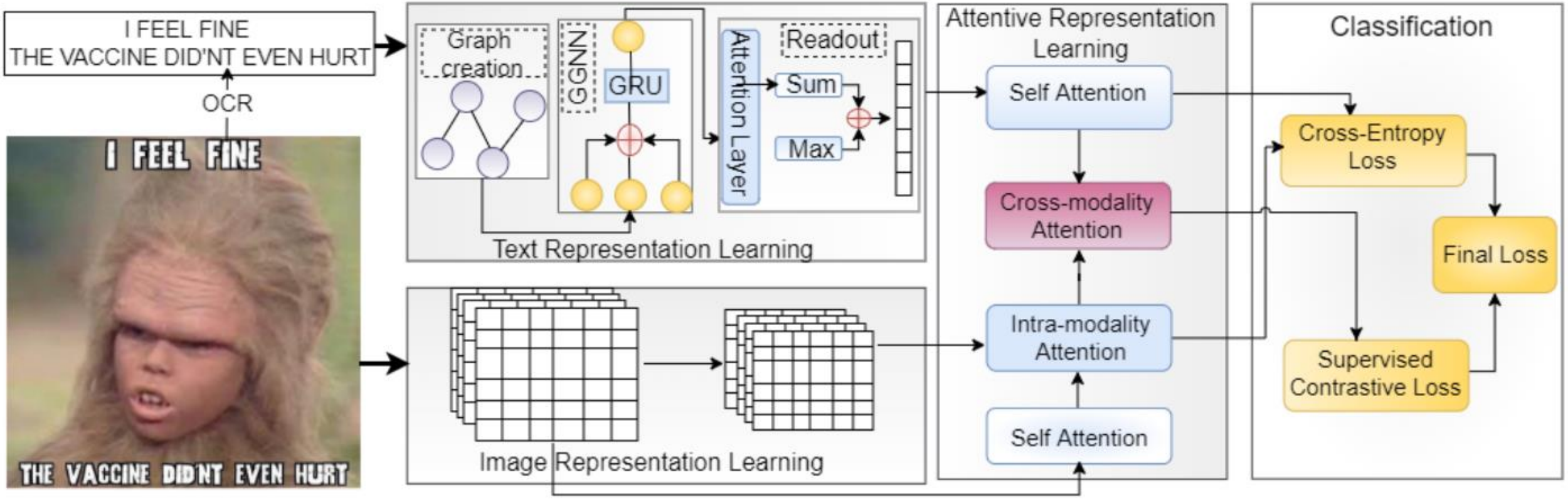


Figure 2: Overall architecture of the proposed multimodal framework.

Method

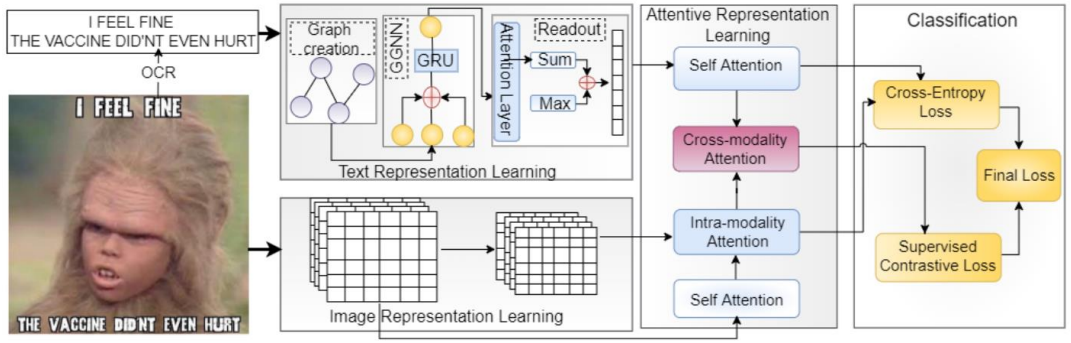
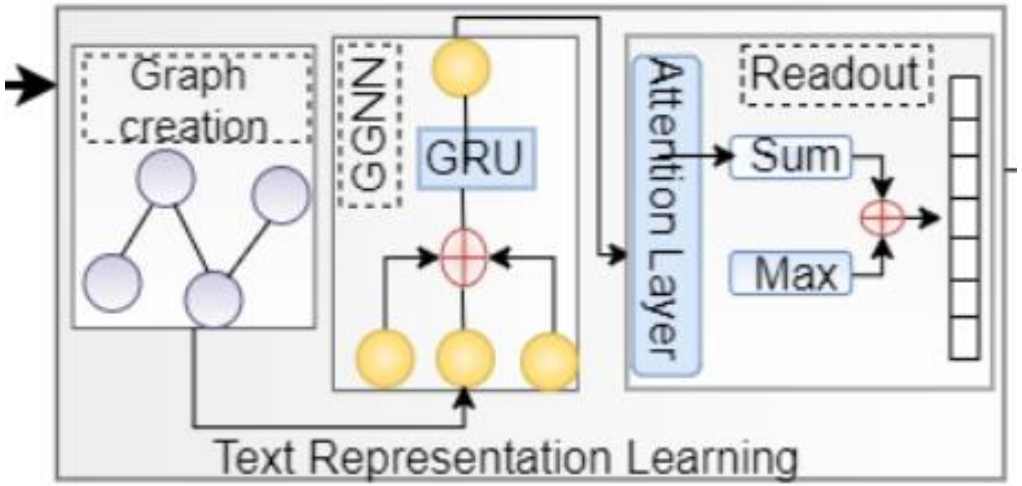


Figure 2: Overall architecture of the proposed multimodal framework.



Text representation learning

graph creation:

$G = (V, E)$ Vertex embeddings $h \in \mathbf{R}^{|V| \times d}$

Word relationship:

$$a^t = Ah^{t-1}W_a, \tag{1}$$

$$z^t = \sigma(W_z a^t + U_z h^{t-1} + b_z), \tag{2}$$

$$r^t = \sigma(W_r a^t + U_r h^{t-1} + b_r), \tag{3}$$

$$h^t = tanh(W_h a^t + U_h(r^t \cdot h^{t-1}) + b_h), \tag{4}$$

$$h_t = h^t \cdot z^t + h^{t-1} \cdot (1 - z^t), \tag{5}$$

Readout Operation:

$$h_v = \sigma(f_1(h_v^t)) \cdot tanh(f_2(h_v^t)), \tag{6}$$

$$h_G = \frac{1}{|V|} \sum_{v \in V} h_v + Maxpooling(h_1, \cdots h_v), \tag{7}$$

Method

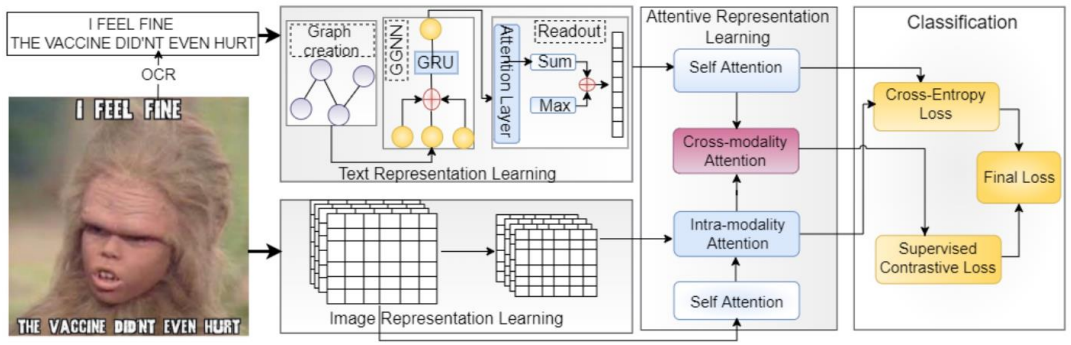


Figure 2: Overall architecture of the proposed multimodal framework.

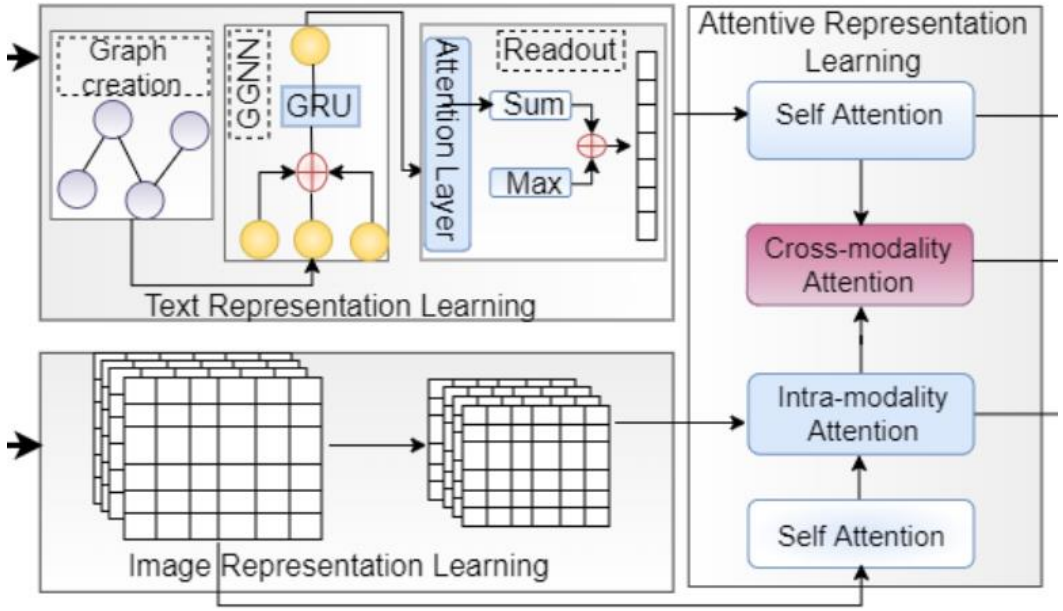


Image representation learning

global f_g and local f_l image features

Attentive representation learning

Single modality-based attention:

$$h_G^{attn} = W_{h_G} \otimes h_G; \tag{8}$$

$$f_l^{attn} = W_{f_l} \otimes f_l; \tag{9}$$

Cross-modality-based attention:

$$F_{Meme}^{V} = (1 + a_v) F_I^{attn} \tag{10}$$

$$F_{Meme}^{T} = (1 + a_t) h_G^{attn} \tag{11}$$

$$F_{Meme} = W_F \otimes [F_{Meme}^{V}, F_{Meme}^{T}] \tag{12}$$

Method

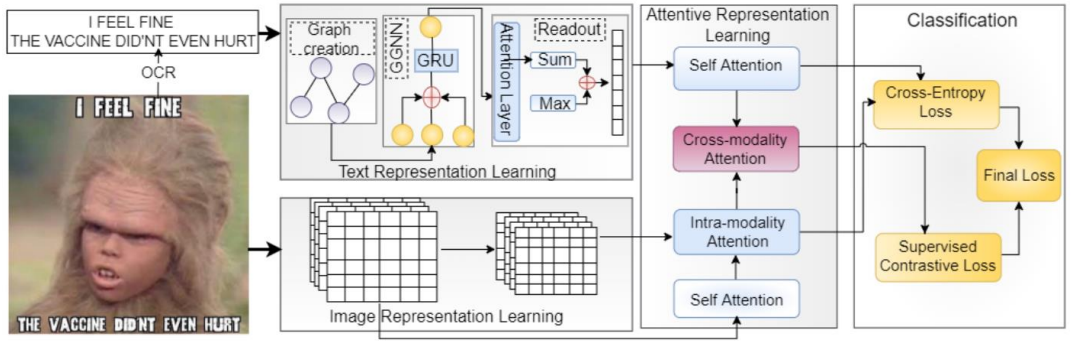
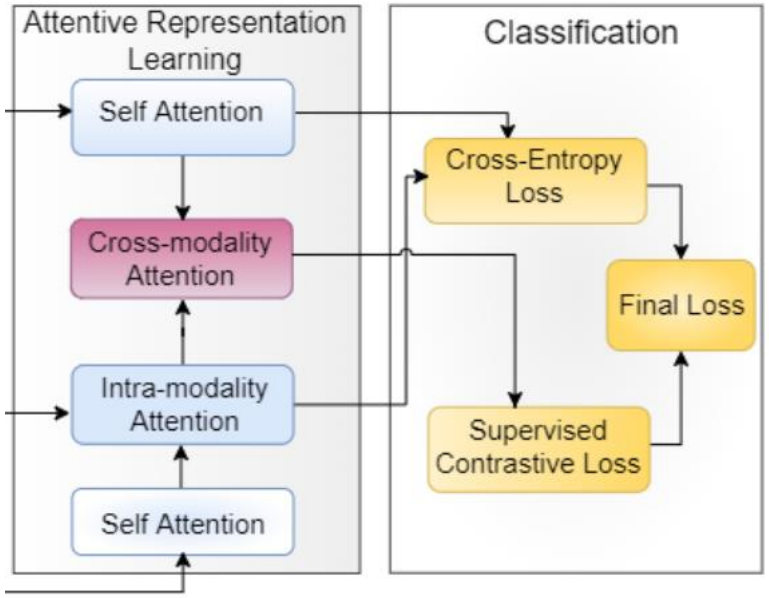


Figure 2: Overall architecture of the proposed multimodal framework.



Classification

$$L_{SCL} = \sum_{i=1}^{N} \frac{-1}{N_{\hat{y}_i} - 1} \sum_{j=1}^{N} 1_{i \neq j} \cdot 1_{\hat{y}_i = \hat{y}_j} \cdot \log\left(\frac{exp(z_i) \cdot (z_j)/\tau)}{\sum_{k=1}^{N} exp(z_i) \cdot (z_k)/\tau)} \right) \tag{13}$$

$$L_{CE} = -\sum_{c=1}^{c} y \log(\hat{y}) \tag{14}$$

$$L = (1 - \lambda) L_{CE} + \lambda L_{SCL} \tag{15}$$

Experiments

Table 2: Dataset Statistics

Data	No. of Pro-Vaccine	No. of Vaccine critical	No. of Neutral	Total
Full Dataset	3983	3441	2820	10244
Timeline 1 (T1)	452	1679	1027	3158
Timeline 2 (T2)	1040	747	1062	2849
Timeline 3 (T3)	2491	1015	731	4237

Experiments

Table 3: Comparison: Proposed framework v/s the baselines. * shows that our proposed framework obtained a significant ($p < 0.05$) performance improvement over the second best approach (underlined) under Mann–Whitney U test.

Type	Model	F1-Score	Precision	Recall
Text only	LSTM	68.48	69.22	68.69
	GRU	68.56	68.73	68.73
	BERT	72.69	72.81	75.75
	TextGCN	73.60	73.30	74.50
	BertGCN	74.10	74.00	74.80
Image only	DenseNet	61.42	63.68	62.88
	ResNet	58.99	63.62	61.36
	VGGNet	58.57	61.65	60.60
Multimodal	ViLBERT	77.23	76.73	76.27
	VisualBERT	79.33	78.84	78.25
	MMBT	78.97	78.61	78.13
	DisMultiHate	80.10	80.35	79.10
	MVAE	80.67	81.00	79.58
	EANN	80.78	81.13	79.69
	MOMENTA	80.07	81.22	81.02
	att-RNN	81.15	81.48	80.04
	DGExplain	81.50	81.90	80.00
	SeTa-Attn	81.65	82.36	80.96
	Proposed	84.20*	85.00*	83.42*

Experiments

Timeline\Models		T1			T2			T3		
		F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall
T1	BertGCN	71.20	71.50	71.40	66.10	67.00	67.10	67.50	67.40	67.30
	DensNet	55.07	55.23	58.78	51.02	52.94	55.39	52.09	53.30	54.28
	SeTa-Attn	73.29	74.27	73.05	67.53	68.11	70.09	69.16	69.25	69.20
	Proposed	**78.25**	**78.59**	**79.16**	**71.78**	**71.62**	**74.73**	**74.83**	**74.37**	**76.54**
T2	BertGCN	59.90	62.80	59.50	70.10	71.00	71.80	65.20	63.15	60.20
	DensNet	49.84	52.64	54.55	54.32	53.94	57.69	50.18	51.94	53.68
	SeTa-Attn	62.68	63.75	63.03	73.47	72.88	72.84	68.15	68.40	68.58
	Proposed	**71.11**	**70.49**	**72.71**	**78.76**	**77.29**	**80.29**	**75.58**	**75.27**	**77.64**
T3	BertGCN	62.60	66.35	62.50	67.50	67.90	68.20	71.20	71.90	71.50
	DensNet	47.68	50.45	54.08	50.64	52.42	57.03	53.80	54.21	58.56
	SeTa-Attn	66.87	64.83	69.51	68.58	69.16	72.91	74.61	73.97	75.27
	Proposed	**71.25**	**71.09**	**71.92**	**75.82**	**75.21**	**77.30**	**79.18**	**78.64**	**78.39**

Experiments

Table 5: Ablation analysis: Proposed framework w/o SCL shows the result of using cross-entropy only as a loss function, i.e., without a supervised contrastive loss (SCL) from the final loss. Proposed w/o ARL shows the results without the attention representation learning (ARL) module from the proposed method. Proposed w/o image ad proposed w/o text represent the results without image and test features in the proposed method. *indicates that the proposed framework obtained a significant ($p < 0.05$) performance improvement over other variants of the proposed method under the Mann–Whitney U test.

Method	F1-Score	Precision	Recall
Proposed	84.20*	85.00*	83.42*
Proposed w/o SCL	82.70	82.86	82.82
Proposed w/o ARL	80.17	79.86	80.16
Proposed w/o image features	78.40	78.51	78.45
Proposed w/o text features	64.10	64.66	65.35

Experiments

		
Ground Truth:Vaccine critical	**Ground Truth**:Vaccine critical	**Ground Label**: Pro-Vaccine
Predicted Label:Vaccine critical	**Predicted Label**: Vaccine critical	**Predicted Label**: Pro-Vaccine

Figure 3: Qualitative analysis: Examples of memes that are correctly predicted by the proposed method.

Experiments

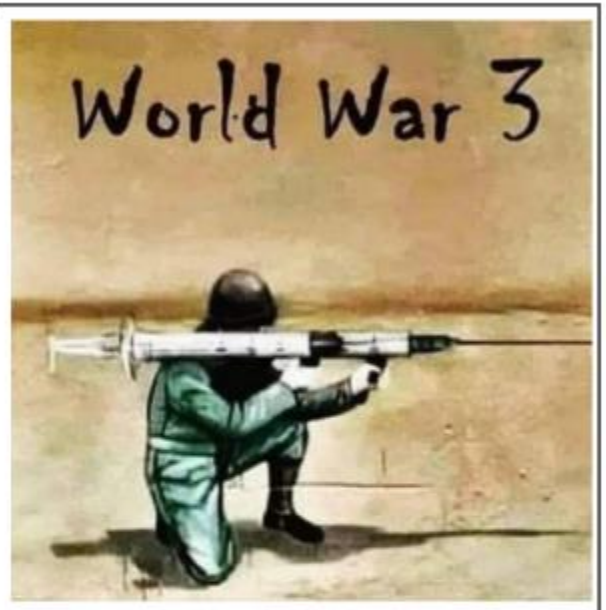
	
Ground Truth: Vaccine critical	**Ground Truth**:Vaccine critical
Predicted Label: Neutral	**Predicted Label**: Neutral

Figure 4: Error analysis: Examples of the memes that are incorrectly predicted by the proposed method.

Thank you!