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

ABSTRACT

Memes can be a useful way to spread information because they are funny, easy to share, and can spread quickly and reach further than other forms. With increased interest in COVID-19 vaccines, vaccination-related memes have grown in number and reach. Memes analysis can be difficult because they use sarcasm and often require contextual understanding. Previous research has shown promising results but could be improved by capturing global and local representations within memes to model contextual information. Further, the limited public availability of annotated vaccine critical memes datasets limit our ability to design computational methods to help design targeted interventions and boost vaccine uptake. To address these gaps, we present VaxMeme, which consists of 10,244 manually labelled memes. With VaxMeme, we propose a new multimodal framework designed to improve the memes' representation by learning the global and local representations of memes. The improved memes' representations are then fed to an attentive representation learning module to capture contextual information for classification using an optimised loss function. Experimental results show that our framework outperformed state-of-the-art methods with an F1-Score of 84.2%. We further analyse the transferability and generalisability of our framework and show that understanding both modalities is important to identify vaccine critical memes on Twitter. Finally, we discuss how understanding memes can be useful in designing shareable vaccination promotion, myth debunking memes and monitoring their uptake on social media platforms.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Vaccine critical memes analysis, Multimodal data and framework

WSDM '23, February 27-March 3, 2023, Singapore, Singapore
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570450

ACM Reference Format:
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 (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570450

1 INTRODUCTION

Memes are typically images with embedded text that use humour and satire, often aimed at encouraging sharing over social media. A meme that incorporates false or misleading information can be weaponised for spreading misinformation and potentially harmful beliefs and attitudes [33]. Vaccine critical information may be associated with potentially harmful attitudes like vaccine hesitancy or behaviours like vaccine refusal, and erode confidence in vaccination campaigns [33, 40]. Evidence from social psychology and population-level studies has shown connections between the information people are exposed to online and the opinions they express [7, 8].

Meme analysis on social media platforms, including Twitter, can help identify emerging threats and guide targeted communication interventions, education, and policymaking [37]. Memes that include vaccine critical information appear to have increased in parallel with the introduction of COVID-19 vaccines [10], and memes on social media are a popular source of vaccine misinformation [40]. Meme analysis is challenging because memes rely on sarcasm and satire, which rely on understanding the context of the image and text. In many cases, the visual and textual content use juxtaposition for humour, and the meme's meaning can be in direct contrast with the text (Figure 1).

Most research on vaccines has focused primarily on using textual content. However, posts with images are common and can alter the text's meaning, which creates a challenge in making sense of the spread of misinformation and the influence of vaccine critical posts. The limited public availability of annotated multimodal datasets limits our ability to investigate new ways of interpreting and classifying social media posts that include vaccine critical memes.

Recent research on meme analysis has focused on tasks like detecting hateful memes [21, 34], misleading information [38] and fake news detection [39]. However, these methods, including the

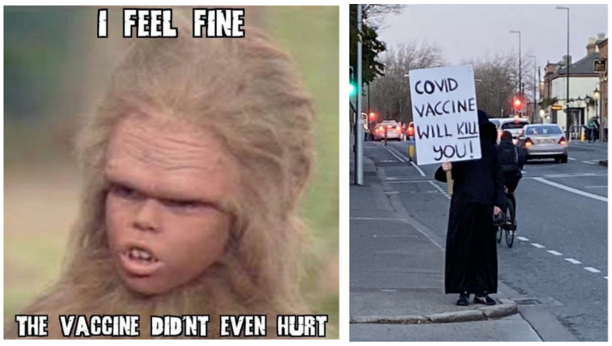


Figure 1: Examples of vaccine critical memes. Note that in a meme shown on the left, an image becomes a humorous way to identify that a meme is vaccine critical, whereas, for a meme on the right, a text suggests that a meme is vaccine critical.

previous state-of-the-art method to identify vaccine critical information on social media [40], used various language models to extract textual features that focus on the locality of words, and they hence lack the long-distance and non-consecutive word interactions required to capture global features. While previous work on identifying vaccine critical information using social media [40] achieved good performance on a multimodal dataset created from Instagram posts, it may not be able to capture global and local representations of both textual and visual content within memes and fails to capture contextual information because (i) the meme's textual and visual content are frequently unrelated and need to be interpreted together; (ii) interpretation of the memes often relies on contextual knowledge, and (iii) methods that use a cross-entropy loss function has limitations that can lead to instability [14].

In this study, we aim to overcome the above limitations by presenting VaxMeme[1], a new multimodal data and framework to identify vaccine critical memes on Twitter. Our contributions are:

- 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% (an absolute increase of 2.6% compared to the best baseline method) and also establish the transferability and generalisability of the proposed framework.

2 RELATED WORK

Social media is a valuable source of information and has been widely used for various tasks like health mention classification [31], Identifying suicide [30] and depression [28], and others [32]. Systematic reviews show the wide range of applications for classifying user-generated content for vaccine hesitancy on social media, such as infectious diseases and outbreaks such as human papillomavirus [43], measles Influenza [4], mining misinformation mining [6].

2.1 Existing datasets

Only two multimodal datasets are used in the previous studies to identify vaccine critical information on social media. The first of them was presented by Wang et al. [40], where authors used Instagram posts with text and visual content collected from January 2016 to October 2019 to identify vaccine critical information on Instagram posts. MMCoVaR [2], a multimodal COVID-19 vaccine focused data repository is the second dataset. MMCoVaR comprises 2,593 articles and 24,184 tweets from February 2020 to March 2021 and is limited to only COVID vaccine related posts. Both mentioned datasets are not publicly available, whereas we make our dataset publicly available for further research.

2.2 Existing methods for vaccine critical posts

2.2.1 Methods for textual data: Majority of research on identifying vaccine critical posts on social media has mainly focused on textual content [40]. Zhang et al. [44] presented three models for analysing public sentiment on HPV vaccines on Twitter using transfer learning. They fine-tuned bidirectional encoder representations from Transformers (BERT) [5], and their results demonstrated the effectiveness of the proposed framework, which also aided in the discovery of vaccine uptake strategies. Recently, Naseem et al. [29] categorised vaccine-related tweeter posts using word representation from the domain-specific context with common knowledge and sentiment data. Their proposed method outperformed several traditional and recent transformer-based pre-trained language models. Previously published architectures, however, only focus on local semantic word representations using a sliding window for textual content. However, long-range and non-consecutive semantic links among feature representation words are required to capture global characteristics. We address this limitation by using a graph-based method to capture both local and global features of textual content.

2.2.2 Methods for multimodal data: Previously research has examined the use of multimodal content for detecting hateful memes [21, 34], misleading information [38], antisemitism [1], and fake news detection [39]. Experiments conducted using unimodal and multimodal in previous studies showed that understanding both modalities is essential for detection. Limited research has explored multimodal data to identify vaccine critical memes on social media. Recently, Wang et al. [40] created a multimodal dataset from Instagram posts and presented a multimodal framework with semantic and task-level attention to identifying vaccine critical information on social media. In contrast, our work jointly learns global and local representations of the textual and visual content of memes, which provide complementary information to improve the identification of vaccine critical memes on Twitter. We suggest that releasing a robustly annotated dataset to the community will support further advances and benchmarking of methods in this space.

2.3 Contrastive learning

Contrastive learning aims to learn an embedding space with positive pairs close together and negative pairs far apart. It is like metric learning, which seeks to learn a distance function in an embedding space [41]. Supervised contrastive loss-based methods have demonstrated considerable success in a variety of tasks, including visual

[1] https://github.com/usmaann/VaxMeme

Table 1: We provide annotation instructions given to annotators. We also show the three most salient words for each label found using sparse additive generative (SAGE). A higher SAGE coefficient indicates prominence within the corpus of that class.

Label	Instructions	Most salient words	SAGE coefficient
Pro-vaccine	A meme (text or image or both) contains a content in favor of vaccines, advising people to get vaccinated, a content about any event or place that is open only for vaccinated people or promoting and selling products with slogans in favor of vaccines.	vaxxed vaccinated get vaccinated	1.73 1.68 1.67
Vaccine critical	A meme (text or image or both) criticises vaccines, contains vaccine misinformation about vaccine side effects, vaccine conspiracy theories, and cases or statistical conclusions against vaccines.	no vaccine for me depopulation vaccine death	1.74 1.72 1.70
Neutral	A meme (text or image or both) reports the events or others' opinions objectively related to vaccines, such as talking about rights of people related to vaccines, or news or statistical charts about vaccines showing no content in favor or against vaccines.	vaccine passport coronavirus outbreak	0.62 0.61 0.50

representation learning [9], few-shot learning [26] and others [12]. However, to our knowledge, no study has used supervised contrastive learning to identify vaccine critical memes on social media. We fill this gap and introduce an optimised loss function that uses supervised contrastive loss to improve the identification of vaccine critical memes on Twitter.

3 DATA

Data collection and annotation: We augmented a publicly available dataset released by Muric et al. [27]. Using the released tweet ids, we collected only those tweets that contained both text and images from October 2020 to April 2021, using the Twitter application programming interface (API). We excluded tweets in languages other than English, and the textual content was automatically extracted using Optical Character Recognition (OCR).

Annotators were given the original image and the OCR text to label the data following the annotation guidelines. Our annotation team includes 8 participants who are fluent in English and hold academic degrees ranging from MSc to PhD. Annotation requires an understanding of a meme's textual and visual components.

Each meme was annotated independently during the annotation process. Where there was a disagreement among the annotators, a third annotator was assigned, and a majority vote decided the label. To ensure consistency, memes were given to annotators in batches. Fleiss' kappa (κ) (coefficient of agreement between annotators) was high (κ = 0.85) in a randomly selected sample of 700 memes.

Annotators were provided with instructions (Table 1) prepared by following a similar study on identifying vaccine critical posts on social media by Wang et al. [40]. Before starting the annotation, annotators were instructed to read the annotation instructions. Following that, more discussions were held to determine whether the annotation instructions were clear to the annotators. Annotators were instructed to select one of the three labels, i.e., pro-vaccine, vaccine-critical, or neutral.

Data analysis: Table 1 shows the linguistic variance across the labels in VaxMeme, obtained by analysing the sparse additive generative (SAGE) model [11] that combines topical and generalised additive models. SAGE implies that by identifying the differentiating words, we can determine the relative relevance of a class to all other classes by comparing word distributions between a target corpus and a reference corpus that use the metric of a log-odds ratio. Because of the additive nature of SAGE, we can determine which words have a significant impact on each label. The cluster of words for the *vaccine-critical* label contains strong negative words such as 'no vaccine for me,' 'depopulation,' and 'vaccine death,' as anticipated. As for the memes labelled as *pro-vaccine*, we observe clear indications of words in support of vaccines like 'vaxxed,' 'vaccinated,' and 'get vaccinated.' Finally, neutral words like 'vaccine passport,' 'coronavirus,' and 'outbreak.' are used for memes labelled as neutral. There is a shift in memes across both labels, as shown by the most salient words in each class.

Data statistics: Based on the above steps, we constructed a new multimodal dataset containing 10,244 memes (Table 2).

For real-time scenarios where the count of memes and differences in content requires computational methods to identify general properties of vaccine critical memes to be useful in deployment, we divided our dataset into 3 timelines (i.e., T1, T2, and T3) based on the months a meme is posted (Table 2).

T1 contains memes from October 2020 to February 2021, when more anti-vaccine memes were posted due to the following: FDA-approved COVID vaccines for emergency use, and various articles and news were published about Pfizer vaccine particles that cause a rare allergic reaction etc. T2 contains memes only from March 2021, where we observed a rise in pro-vaccine memes due to the availability of COVID vaccines. In T3, the number of pro-vaccine memes increased in April 2021, indicating people's support for vaccines increased.

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

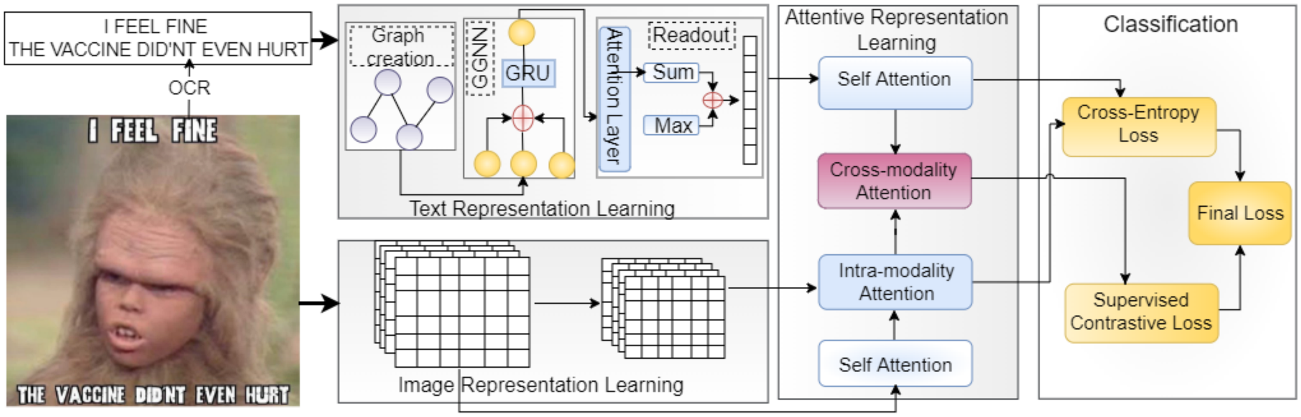


Figure 2: Overall architecture of the proposed multimodal framework.

4 METHODOLOGY

Overview of our framework: Figure 2 shows the proposed multimodal framework. It consists of text representation learning, image representation learning, attentive representation learning, and classification. Below, we describe each module in depth.

4.1 Text representation learning

Our text representation learning module uses a graph neural network-based method to extract global and local text representation. It consists of 3 key steps: graph creation, word relationship, and readout operation.

Graph creation: Following [45], we create the graph where the unique words are represented as vertices, whereas the co-occurrences between words are edges. We represent this graph creation step as $G = (V, E)$ where V is the set of vertices and E is the set of edges. The co-occurrences explain the relationship between words that appear in the graph within a fixed-size sliding window (length 3 by default). Vertex embeddings are initialised with word features, represented by $h \in \mathbf{R}^{|V| \times d}$ where d is the embedding dimension. We propagate and contextualise word feature information during the word interaction by creating individual graphs for every document.

Word relationship: We leverage gated graph neural network [23] to extract the representations of the word nodes on each graph. A node receives information from its neighbours and integrates it into its representation to update. Because the graph layer functions on first-order neighbours, we can stack such layers t times to obtain high-order feature relationships, in which a node can reach another node t hops away. This step of the word relationship is mathematically represented as follows:

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

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

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

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

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

where $A \in \mathbf{R}^{|V|x|V|}$ denotes the adjacency matrix, σ represents the sigmoid function, and trainable weights and biases are represented by U, W, and b. r and z represent the reset and update gates, respectively, to identify the extent to which neighbour information adds to the current embedding of a node.

Readout Operation: The word nodes are combined to form a graph-level representation of the post, which is then used to generate the final prediction after being sufficiently updated. The readout operation [45] is defined as follows:

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

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

f_1 and f_2 denotes 2 multi-layer perceptrons. The former (f_1) works similarly to a soft attention weight, whereas the latter (f_2) is a non-linear feature transformation. We use max-pooling for the graph representation h_G in addition to averaging the weighted word features. The assumption is that each word in the text has a meaning, and the keywords should be used more explicitly.

Above generated graph-based text representation h_G, containing both global and local text information, is fed to self-attention in our attentive representation learning module.

4.2 Image representation learning

Our image representation learning module uses ResNet-50 [15] as the backbone to obtain image features. The global image features $f_g \in R^C$ are obtained from the final adaptive average pooling layer of the ResNet-50 model, where C represents the feature dimension. We obtained the local image features from an intermediate convolution layer and vectorized them to get the C-dimensional features for each of the M image sub-regions: $f_l \in R^{CxM}$.

Above generated image representation, containing both global f_g and local f_l image features, are fed to self-attention and intra-modality attention in the attentive representation learning module.

4.3 Attentive representation learning

Single modality-based attention: We utilise the self-attention mechanism to emphasise the important content of both textual and visual features. For textual features, we feed the graph-based text representation h_G, containing both global and local text information to a self-attention to emphasise the important textual features h_G^{attn}.

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

where W_{h_G} is a learnable parameter, and $\otimes$ represents the outer product of a matrix.

For visual features, we feed local image features (f_l) to the attention module to capture local image features f_l^{attn}.

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

where W_{f_l} is a learnable parameter, and $\otimes$ denotes the outer product of a matrix.

Following that, we use an intramodality attention module to integrate the self-attended image local features (f_l) with the global image features (f_g) to capture image features (F_I^{attn}). This stage aims to integrate the local image descriptions with the meme's global semantics. Overall, the meme's semantics are captured by local and global features, considering the meme's background context.

Cross-modality-based attention: We observed that text embedded in memes is more relevant for some memes, whereas, for others, image plays a vital role in identifying vaccine critical memes (Figure 1). To address this, our cross-modality-based attention uses an attention method to integrate the representations from both the textual and the visual modalities. Inspired by previous work by Gu et al. [13], we create our cross-modality-based attention fusion with two major steps: generation of modality attention and concatenating weighted features. The attention scores $[a_v, a_t]$ for the two modalities are generated in the first step using a sequence of dense layers followed by a softmax layer.

In the second step, we combine the original unimodal features by weighing them according to their respective attention scores. For better gradient flow, we also employ residual connections.

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

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

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

where, W_F is a learnable parameter, and F_{Meme} is the final representation. We feed the final meme representation obtained from the cross-modality attention module into a fully-connected layer for the final classification.

4.4 Classification

Supervised contrastive Loss: Motivated by the learning approach that humans use when provided with few examples, we focus on finding similarities between the examples of each class and compare them with examples from other classes. We propose using a contrastive learning loss that captures similarities between examples from the same class and contrasts them with examples from other classes in the embedding space to leverage the label information better. We define the contrastive loss as:

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

Where $N_{\hat{y}_i}$ is the total number of memes in the minibatch that have the same label, $\hat{y}_i$, as the anchor, i, the dot product $(\cdot)$ denotes similarity score between positive and negative examples of a vector z, and τ is a scalar temperature parameter which controls the effect of the hard negatives in the training process.

Cross-entropy Loss: We fed the representations (h_G^{attn} and F_I^{attn}) from the single modality-based attention to the cross-entropy loss, which is defined as:

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

where y represents the true class label and C represents the total number of classes in the dataset.

Final Loss: Our final loss function is a weighted average of cross-entropy and the supervised contrastive loss, as given in equation (15). The canonical definition of the multi-class cross-entropy (CE) loss is given in equation (14). The supervised contrastive loss (SCL) is given in equation (13). We optimise the following final loss when training our model:

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

where λ controls how much each loss term contributes.

5 EXPERIMENTS

5.1 Experimental settings

We used similar experimental settings for all experiments and 10-fold cross-validation for consistency. F1-Score, Recall, and Precision scores are used to examine the performance of our proposed framework, and the average of all folds is reported. Our proposed framework can be trained end-to-end with backpropagation, and we performed gradient-based optimization using the Adam update rule with a learning rate of 0.001. We used the base version of pretrained language models using the HuggingFace library. Variable-length posts are padded and trained for 50 epochs.

5.2 Baselines

5.2.1 *Unimodal Models:*

For the unimodal-based baselines, we used text-only and image-only-based methods, commonly used in similar previous studies (explained in section 2), to compare the performance. We used the following methods:

- Unimodal - Text only: For the text-only based unimodal, we used long-short term memory (LSTM) [16], gated recurrent unit (GRU) [3], and a BERT [5], a state-of-the-art pretrained transformer-based language model. We also compared the results of our method with state-of-the-art graph-based methods like TextGCN [42] and BertGCN [24].
- Unimodal - Image only: For the image-based unimodal baselines, we used DenseNet [17], VGGNet [36] and ResNet [15].

5.2.2 ***Multimodal Models:*** Multimodal-based methods have shown good performance over unimodal-based methods over a wide range of multimodal tasks. We used the following multimodal-based methods, which have been widely used in previous studies.

We experimented with a multimodal method trained using a multimodal objective. In particular, we used Vision and Language BERT (ViLBERT) [25], VisualBERT [22], Multimodal Bitransformer (MMBT) [20], Event adversarial neural networks (EANN) [39], Multimodal variational autoencoder (MVAE) [19], Recurrent Neural Network with an attention mechanism (att-RNN) [18], MultimOdal framework for detecting harmful memes and their targets (MOMENTA) [34], and DisMultiHate [21]. In addition, we also compared our results with Duo-generative explainable (DGExplain) [35], a generative method for identifying multimodal COVID-19 misinformation that evaluates the cross-modal relationship between visual and textual content in multimodal news content and SeTa-Attn [40], a multimodal deep neural network with semantic and task-level attention for detecting vaccine critical posts on social media.

6 RESULTS

6.1 Comparison with Baselines

Overall comparison: Table 3 presents the overall performance of our multimodal framework when compared to the state-of-the-art methods. The performance of using text-only methods is better than image-only methods; this is as expected, given that text contains more explicit information than visual information in posts. Further, BERT outperforms GRU and LSTM, which is expected because BERT can better capture contextual representation than LSTM and GRU. We also observed that graph-based methods perform better when compared to other text-only baselines. We attribute this increase to the possibility of better capturing global and local text representation of a user post. In addition, we also note that the performance of both (text only and image only) models is low, i.e., no higher than an F1-Score of 74.10%. As a result, textual and visual data must reflect the distinct characteristics of a meme. Following that, we show how combining texts and images helps improve performance when identifying vaccine misinformation in memes.

We observed higher performance in multimodal methods compared to only text-only and visual-only models (Table 3). Further, the state-of-the-art multimodal methods, i.e., ViLBERT, VisualBERT, and NMBT, that are designed for classifying vision-and-language tasks perform poorly compared to other multimodal methods, i.e., EANN, MVAE, att-RNN, MOMENTA, DisMultiHate and SeTa-Attn that are designed for the identification of specific tasks such as rumour detection, fake news detection, hateful memes detection and vaccine misinformation detection on social media. VisualBERT achieved an F1-Score of 79.33%, whereas SeTa-Attn, designed to detect vaccine misinformation using multimodal data, achieved an F1-Score of 81.65%.

We observed that the proposed method outperformed all tested methods, both unimodal and multimodal modal, and achieved an F1-Score of 84.20%, an absolute increase of 2.55% when compared to SeTa-Attn (best baseline), which is designed to capture vaccine misinformation using multimodal data. We also observe that the performance of other multimodal methods (i.e., MVAE, EANN, SDM, DGExplan, and MTAN) that are designed for other tasks are less desirable in detecting vaccine critical memes on Twitter. The reason is that these methods focus on local representations of visual and textual content within memes and ignore global representations, i.e., long-distance relations, resulting in low performance.

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*

Transferability of proposed method: Compared to the best performing baselines from each modality, our proposed method outperforms the baselines and shows the smallest degradation in performance when tested on data from different timelines, i.e., T1, T2, and T3 (Table 4). To evaluate the transferability, we first train and test all models on each timeline separately, then train on one timeline and test the other two timelines. We demonstrate that the proposed method outperformed the best-baseline method with an absolute increase ranging from 3.44% to 4.57% in terms of F1-Score when trained and tested on the same timeline. Further, when tested on other timelines, the proposed method achieved an absolute increase of 4.65% and 5.84% F1-Score when models were trained on T1 and tested on T2 and T3 timelines, respectively. For T2, the proposed method outperformed SeTa-Attn with an absolute increase of 9.39% and 8.04% F1-Score when tested on T1 and T3, respectively. Finally, for T3, the proposed method outperformed SeTa-Attn by 4.38% and 6.27% F1-Score when tested on T1 and T2, respectively. This transferability shows the robustness and generalisability of the proposed method on new unseen results.

6.2 Analysis

Ablation analysis: An ablation analysis shows that each of the new components we included in the proposed multimodal framework contributed to the overall performance (Table 5). The F1-Score

Table 4: Transferability of the best-performing baselines from each modality (Text only, Image only, Multimodal) and the proposed framework on T1, T2, and T3. The models are trained using the timelines in the row and tested using the timelines in the column. Bold indicates the best transferable results for each timeline.

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**

dropped (from 84.20% to 80.17%) when we removed the attentive representation layer (ARL) from our framework. Similarly, performance dropped to an F1-Score of 82.70% when we removed contrastive loss (SCL) from our loss function. Finally, removing the image and text features from the proposed framework resulted in F1-Score dropping to 78.32% and 64.40%, respectively. Hence, we deduce that the strengths of our framework lie in integrating all modules that contribute to increased performance.

Generalisability test: A generalisability test shown on HarmC [34] memes dataset shows that the proposed framework outperformed the best results of 82.80% F1-Score reported in [34] and Seta-Attn with an F1-Score of 83.50% by an absolute increase of 3.50% and 2.80% respectively. Our generalisability test concludes that our framework is generalisable and outperforms the state-of-the-art and Seta-Attn in other (i.e., harmful) meme classification tasks.

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.

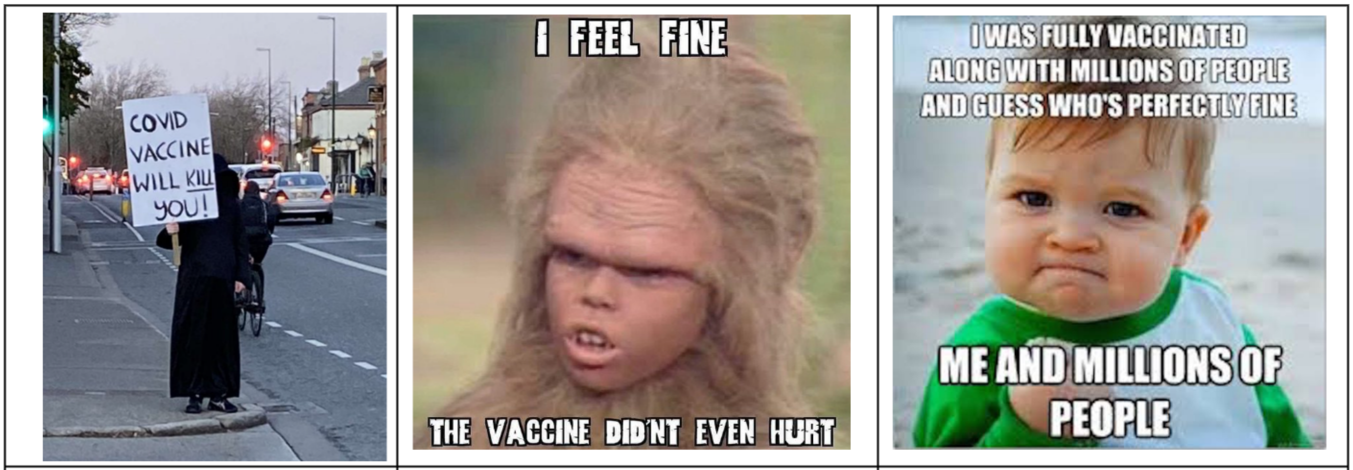
Method	F1-Score	Precision	Recall
Proposed	84.20*	85.00*	83.42*
Proposed w/o SCL	82.70	82.86	82.82
Proposed w/o ARL	80.17	79.86	80.16
Proposed w/o image features	78.40	78.51	78.45
Proposed w/o text features	64.10	64.66	65.35

Qualitative analysis: Figure 3 shows examples of memes correctly predicted by our framework. We can observe that only capturing one modality of data may result in a wrong prediction as it does not capture the underlying context of a meme. For example, in the second meme in Figure 3, if we only focus on the text, then the prediction would be wrong as the context of the meme changes with the visual content, which shows that it is important to capture both modalities of data to identify the memes with misinformation correctly. Based on these examples, we can conclude that our method, which captures both modalities with both global and local features and captures important features due to our attention representation learning mode, outperforms both unimodal and multimodal baselines, including the current state-of-the-art method.

Error analysis: We discuss some of the cases where our method failed: (i) posts where there is insufficient information, i.e., both image and text do not contain words or user opinions against vaccines, (ii) the OCR failed to detect the words; and (iii) posts that are challenging to make correct predictions without having additional domain knowledge. Examples of memes that are incorrectly predicted by the proposed framework are shown in Figure 4.

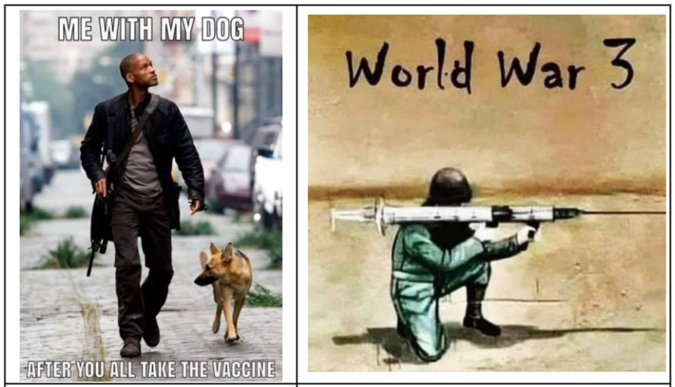
7 PRACTICAL IMPLICATIONS

Memes are an effective way to spread vaccine critical information on social media. Social media platforms and public health organisations need to quickly identify social media posts that are vaccine critical or may spread misinformation so that they can act. This may need to be done for individual posts rather than for users; actions from social media platforms might include downranking and flagging posts for their content, and actions from public health organisations might include targeted communication interventions and debunking emerging myths. For memes specifically, understanding how memes become popular and spread may also be useful for


COVID VACCINE WILL KILL YOU!
I FEEL FINE
THE VACCINE DIDN'T EVEN HURT
I WAS FULLY VACCINATED ALONG WITH MILLIONS OF PEOPLE AND GUESS WHO'S PERFECTLY FINE
ME AND MILLIONS OF PEOPLE

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.



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.

designing shareable vaccination promotion and myth-debunking memes and monitoring their uptake on social media platforms.

The proposed method improves over existing approaches for classifying memes and illustrates the importance of interpreting the text and image together and in context because memes often use sarcasm and juxtaposition to mean something other than what the literal interpretation of the text or image might suggests. We suggest that releasing a robustly annotated dataset to the community can facilitate benchmarking of methods and lead to advancements in this important space. While this study was focused on vaccines, we expect the proposed method would also be useful for other scenarios in which memes are used to spread incorrect or misleading information. Finally, the proposed multimodal framework cannot be immediately applied to realistic scenarios. It would be a component of an action plan that calls for collecting more and better-labelled data, conducting trials, and receiving regulatory approval for implementing policy actions against misinformation.

8 ETHICAL CONSIDERATIONS

No human ethics approval was required to analyse data in this study because we annotated data that were published and made available for research purposes, and no participants were recruited. The annotated data we release include de-identified publicly available posts from Twitter, where users well understand public access and there is no expectation of privacy.

9 CONCLUSION

In this work, we study the problem of identifying vaccine critical memes on Twitter. Our contribution is constructing and releasing a new annotated multimodal dataset with 10,244 memes collected from Twitter. In addition, we presented a new multimodal framework that learns global and local visual and textual content within memes. The improved memes' representations are then fed to an attentive representation learning module to capture contextual information for classification using an optimised loss function. Our experimental results demonstrated that our multimodal framework outperformed the state-of-the-art methods. We further demonstrated the transferability and generalisability of the proposed method and showed that understanding both modalities is important to identifying vaccine critical memes on Twitter. We hope that releasing a new annotated multimodal dataset to the community will support further advances and benchmarking of computational methods in this important field.

ACKNOWLEDGMENTS

Authors acknowledge the assistance of Daihua, Mengbo, Mengfei, Anran, and Pengyu in the labelling and initial experimental setup.

REFERENCES

[1] Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. In *13th ACM Web Science Conference 2021*. 148–157.

[2] Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 31–38.

[3] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[4] Michael S Deiner, Cherie Fathy, Jessica Kim, Katherine Niemeyer, David Ramirez, Sarah F Ackley, Fengchen Liu, Thomas M Lietman, and Travis C Porco. 2019. Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health informatics journal* 25, 3 (2019), 1116–1132.

[5] 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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel. 2020. Cascade-LSTM: A tree-structured neural classifier for detecting misinformation cascades. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2666–2676.

[7] Adam G Dunn, Julie Leask, Xujuan Zhou, Kenneth D Mandl, and Enrico Coiera. 2015. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *Journal of medical Internet research* 17, 6 (2015), e4343.

[8] Adam G Dunn, Didi Surian, Julie Leask, Aditi Dey, Kenneth D Mandl, and Enrico Coiera. 2017. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine* 35, 23 (2017), 3033–3040.

[9] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9588–9597.

[10] Marta Dynel. 2021. COVID-19 memes going viral: On the multiple multimodal voices behind face masks. *Discourse & Society* 32, 2 (2021), 175–195.

[11] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 1041–1048.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6894–6910.

[13] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Hybrid Attention based Multimodal Network for Spoken Language Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2379–2390.

[14] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*.

[15] 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.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[17] 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.

[18] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[19] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[20] 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).

[21] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5138–5147.

[22] 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).

[23] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* (2015).

[24] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive Text Classification by Combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1456–1462.

[25] 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).

[26] Orchid Majumder, Avinash Ravichandran, Subhransu Maji, Alessandro Achille, Marzia Polito, and Stefano Soatto. 2021. Supervised Momentum Contrastive Learning for Few-Shot Classification. *arXiv preprint arXiv:2101.11058* (2021).

[27] Goran Muric, Yusong Wu, and Emilio Ferrara. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health Surveill* 7, 11 (17 Nov 2021), e30642. https://doi.org/10.2196/30642

[28] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*. 2563–2572.

[29] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive GRU. *arXiv preprint arXiv:2106.09589* (2021).

[30] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022. Hybrid text representation for explainable suicide risk identification on social media. *IEEE transactions on computational social systems* (2022).

[31] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference 2022*. 2573–2581.

[32] Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam Dunn. 2022. Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. 22–31.

[33] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2783–2796.

[34] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4439–4455.

[35] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A Duo-generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. In *Proceedings of the ACM Web Conference 2022*. 3623–3631.

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

[37] Maryke S Steffens, Adam G Dunn, Julie Leask, and Kerrie E Wiley. 2020. Using social media for vaccination promotion: Practices and challenges. *DIGITAL HEALTH* 6 (2020), 2055207620970785.

[38] Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 659–662.

[39] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[40] Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. *IEEE Journal of Biomedical and Health Informatics* 25, 6 (2020), 2193–2203.

[41] Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research* 10, 2 (2009).

[42] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7370–7377.

[43] Hansi Zhang, Christopher Wheldon, Adam G Dunn, Cui Tao, Jinhai Huo, Rui Zhang, Mattia Prosperi, Yi Guo, and Jiang Bian. 2020. Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. *Journal of the American Medical Informatics Association* 27, 2 (2020), 225–235.

[44] Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. 2020. Sentiment Analysis Methods for HPV Vaccines Tweets Based on Transfer Learning. In *Healthcare*, Vol. 8. Multidisciplinary Digital Publishing Institute, 307.

[45] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. *arXiv preprint arXiv:2004.13826* (2020).