Vaccine Misinformation Detection in X Using Cooperative Multimodal Framework

Usman Naseem
usman.naseem@mq.edu.au
School of Computing, Macquarie University
Sydney, Australia

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

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

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

Abstract

Identifying social media posts that spread vaccine misinformation can inform emerging public health risks and aid in designing effective communication interventions. Existing studies, while promising, often rely on single user posts, potentially leading to flawed conclusions. This highlights the necessity to model users' historical posts for a comprehensive understanding of their stance towards vaccines. However, users' historical posts may contain a diverse range of content that adds noise and leads to low performance. To address this gap, in this study, we present VaxMine, a cooperative multi-agent reinforcement learning method that automatically selects relevant textual and visual content from a user's posts, reducing noise. To evaluate the performance of the proposed method, we create and release a new dataset of 2,072 users with historical posts due to the unavailability of publicly available datasets[1]. The experimental results show that our approach outperforms state-of-the-art methods with an F1-Score of 0.94 (an absolute increase of 13%), demonstrating that extracting relevant content from users' historical posts and understanding both modalities are essential to detecting anti-vaccine users on social media. We further analyze the robustness and generalizability of VaxMine, showing that extracting relevant textual and visual content from a user's posts improves performance. We conclude with a discussion of the practical implications of our study by explaining how computational methods used in surveillance can benefit from our work, with flow-on effects on the design of health communication interventions to counter vaccine misinformation on social media.

CCS Concepts

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

Keywords

Vaccine Misinformation, Multimodal Posts, Cooperative Learning

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

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681422

ACM Reference Format:
Usman Naseem, Adam G. Dunn, Matloob Khushi, and Jinman Kim. 2024. Vaccine Misinformation Detection in X Using Cooperative Multimodal Framework. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3664647.3681422

1 Introduction

Vaccines are an effective means of preventing and controlling infectious diseases. While access to medical care remains a barrier to vaccine coverage, the dissemination of vaccine misinformation can contribute to the risk of outbreaks [22] and impact the capacity to develop vaccines for diseases such as human papillomavirus (HPV) [43] or COVID-19 [23]. Vaccine misinformation can contribute to harmful attitudes such as vaccine refusal and diminish trust in vaccination campaigns [39].

Vaccine misinformation, such as misleading information about effectiveness, emotional narratives about safety, and conspiracy theories, can negatively impact global public health and undermine trust in vaccination efforts [23]. By designing computational methods to detect vaccine critical users on social media automatically, public health organizations and social media platforms can improve how they target communication and education interventions to where and when they are most needed [36].

Analyzing multimodal content on social media platforms, including X (formerly known as Twitter), can help identify emerging threats and can then be used to develop more effective communication strategies and identify key areas of concern that need to be addressed to improve vaccine uptake [22, 36]. The prevalence of multimodal posts on social media containing vaccine misinformation seems to have risen alongside the rollout of COVID-19 vaccines [13] and has become a prominent source of vaccine misinformation [39].

Recent research on the representation of vaccines on social media demonstrated that methods that incorporate both textual and visual information are potentially useful compared to those that use only textual data [4, 39]. Nonetheless, the restricted access to labeled multimodal datasets hinders our capacity to explore novel methods for detecting vaccine misinformation.

Recently, several attention-based approaches have been developed for detecting vaccine misinformation. For instance, Wang et al. [39] introduced a multimodal approach incorporating semantic and task-level attention to highlight key elements in multimodal posts

Figure 1: Examples of the relevant (post #1, post #3, and post #4) and irrelevant (post #2) posts from historical postings of a user who posts vaccine misinformation. User posts are chronologically ordered from left to right.

for detecting vaccine misinformation on social media. Similarly, Shang et al. [33] proposed the Duo-generative explainable (DG-Explain) method, which assesses the interrelation between visual and textual content in multimodal COVID-19 misinformation. In another study by Gao et al. [15] presented multi-agent SelectNet (MASN), a reinforcement learning-based method, leveraging pre-trained BERT and ResNet models to extract text and visual features. MASN utilizes opinion-word and visual-region selectors to enhance personality classification collaboratively, outperforming strong unimodal and multimodal baselines. However, using previous methods focused on encoding users' historical multimodal posts adds noise and leads to low performance. This is because a user's historical posts contain a variety of multimodal content, i.e., vaccine relevant and irrelevant posts (see Figure 1).

In this study, we address the aforementioned limitations by introducing MM-Vax, a novel multimodal dataset of 2,072 X users. We also present VaxMine, which employs a multi-agent reinforcement learning (MARL) approach using two policy gradient agents to concurrently select texts and visuals, and assesses the effectiveness of these joint actions based on classification performance.

To tackle the issue of evaluating each selector's impact in collaborative settings, where combined selections yield global rewards, VaxMine utilizes a unique cooperative inverse operation multi-agent policy gradient approach. This method employs an actor-critic framework with differentiated advantages, where each actor (i.e., text or visual selector) is trained using a specific gradient estimated by a critic. Our approach includes two main components: First, we use centralized training with decentralized execution, where the critic is involved only during the training phase while the actor manages the execution. Second, we incorporate an "inverse operation" technique, providing each agent with a shaped reward reflecting the difference between their current global reward and the reward from opposing action. Our contributions are summarized as follows:

- We construct and release a new multimodal dataset of 2,072 users to detect vaccine misinformation on X.
- We present VaxMine, a cooperative multi-agent reinforcement learning (MARL) based approach where text and visual selectors cooperatively extract the vaccine relevant content to identify vaccine misinformation.
- We show that the performance increases from selecting vaccine relevant content from a user's historical posts and VaxMine outperforms SOTA baselines with an F1-Score of 94% and also establishes the generalizability of VaxMine.

2 Related Work

2.1 Existing Datasets

We know of two multimodal datasets that have been developed for identifying vaccine misinformation using social media. Wang et al. [39] gathered 31,282 Instagram posts, each containing both text and visuals, collected between January 2016 and October 2019. The Multimodal COVID-19 Vaccine Focused Data Repository (MMCoVaR) [5] includes 24,184 X posts with both text and visuals, gathered from February 2020 to March 2021. Unlike these datasets, which are not publicly accessible, we have released our dataset available for further research.

2.2 Existing Methods

Prior methods using textual data: Most research on detecting vaccine misinformation has primarily focused on textual posts and employed traditional machine learning techniques such as SVM [2] and hierarchical SVMs [12]. More recent studies [29] have used bidirectional encoder representations from Transformers (BERT) [9] and its variants to encode user posts on social media.

Other text-based studies [32, 44] that classify a user use different methods to extract textual features and process them sequentially. For instance, Zogan et al. [44] presented DepressionNet, which uses a hybrid extractive and abstractive summarization method to summarize historical posts of a user and then processes them sequentially to identify depression on X. Sawhney et al. [32] used a longformer to extract textual features of all user's posts and sequentially process the historical user posts using BiLSTM to identify suicide risk on social media.

Prior methods using multimodal data: Previous research has explored the application of multimodal posts for identifying various tasks, including fake news [11, 21, 30, 38], Sentiment and Emotion [35], COVID misinformation [33], suicide [3], and depression [6, 42]. Limited research has explored multimodal content regarding social media's role in identifying vaccine critical users. Wang et al. [39] introduced a multimodal deep neural network known as Seta-Attn, which incorporates semantic and task-level attention mechanisms for detecting vaccine misinformation on social media. Previous studies using both unimodal and multimodal approaches demonstrated that comprehending both modalities is crucial for accurately understanding users' opinions.

Recently, researchers also focused on using historical multimodal posts of a social media user for the identification of suicide [3] and depression [6] tasks using different methods to extract textual

Table 1: Annotation instructions with top 3 most salient words and the SAGE coefficient. Higher SAGE coefficients indicate salience in the data of that label.

Label	Instructions	Most salient words	SAGE coefficient
Misinformation	User posts (whether text, visuals, or both) feature vaccine misinformation, criticism of vaccines, vaccine-related conspiracy theories, and cases or statistical arguments opposing vaccines.	depopulation	1.71
		novaccine	1.70
		vaccinedeath	1.69
Otherwise	User either posts in favor of the vaccine or reports events or other perspectives related to vaccines in an objective manner. No content (text or visual) against the vaccine.	amid	0.32
		coronavirus	0.31
		outbreak	0.30

features of all historical posts of a user and processing them sequentially. For instance, Cao et al. [3] presented SDM where they used FastText embeddings of a user's posts and processed them sequentially using an LSTM to capture the temporal dependencies in the user's history for suicide risk identification on social media. Cheng and Chen [6] presented MTAN, using BERT and inceptionResNet, to obtain historical textual and visual content features and process them sequentially for depression detection. Gao et al. [15] introduced the multi-agent SelectNet (MASN), a reinforcement learning-based approach that employs pretrained BERT and ResNet models to extract text and visual features. MASN employs opinion-word and visual-region selectors to collaboratively improve personality classification, surpassing both unimodal and multimodal baseline methods.

The above-mentioned studies explored various encoding, attention strategies, and reinforcement learning techniques. However, these methods alone are insufficient for selecting relevant content, as they may add noise and lead to low performance when applied to the variety of content posted by a user. This is because a user's historical posts contain a variety of multimodal content, including vaccine-relevant and irrelevant posts. We hypothesize that accurately selecting vaccine-relevant content from the variety of posts available can enhance the performance of computational methods in detecting vaccine misinformation on social media.

3 Data

Data selection and collection: We expanded a publicly accessible dataset of X users, provided by Muric et al. [28]. By utilizing the post IDs, we gathered posts with both text and visuals through the X API and organized the collected historical posts by user.

Filtration: We start the filtration step, which includes excluding users with a posting history of fewer than 2 posts or who post in a language other than English. We adopted Optical Character Recognition (OCR) to extract the textual content embedded in the visuals. Annotators were presented with the original visual, the OCR text, and historical posts from a user to annotate the data as per the given annotation instructions.

Annotation: Annotation team comprised 8 members, including men and women, fluent in English. The team members held degrees ranging from MSc to PhD, including expert researchers in NLP and CV. The annotation process involved interpreting both textual and visual content posted by users.

Each user was annotated independently during the annotation. We instructed annotators to label a user with one of two labels (vaccine misinformation or otherwise). Where annotators disagreed, a new annotator was assigned, and the label was assigned by majority vote. Annotators were provided with the posts in batches to ensure consistency. In a random sample of 600 users, a coefficient of agreement between annotators, i.e., Fleiss' kappa (κ) was high (κ = 0.86).

Instructions: Annotators received guidelines (see Table 1) based on a previous study by Wang et al. [39]. Prior to beginning the annotation, annotators were asked to review these instructions. Additional discussions were conducted to ensure that annotators comprehended the guidelines. Specifically, annotators were directed to choose between two labels: "misinformation" or "otherwise". Posts that criticize vaccines, contain vaccine misinformation or conspiracy theories, or present negative cases or statistics about vaccines were categorized as "misinformation". In contrast, a user who reports the events or talks about vaccines objectively and contains no vaccine related misinformation posts in the entire user post history is labeled as "otherwise". In both cases, a user may contain irrelevant posts (Figure 1). If annotators were unsure about a user, they were instructed to remove those users.

Data analysis: By employing Sparse Additive Generative (SAGE) [14], we examine the linguistic variation among labels (see Table 1). SAGE indicates that by identifying key distinguishing words, we can evaluate the relative significance of a class by comparing word distributions between a target corpus and a reference corpus using a log-odds ratio metric. Due to SAGE's additive properties, we can identify which words have a significant effect on each label. For the "misinformation" label, the word cluster includes strongly negative terms such as 'depopulation', 'novaccine', and 'vaccinedeath', as expected. Conversely, the "otherwise" label features clear instances of neutral language, like 'amid', 'coronavirus', and 'outbreak'. The distinct words in each label indicate a shift in users' language across both categories.

Data statistics: Following these steps, we created a new multimodal dataset comprising 2,072 X users (28% misinformation and

Table 2: Dataset Statistics

Total No. of users	2,072
Total No. of posts	30,385
Avg. No. of posts per user	15
Max. No. of posts per user	542
Avg. length of posts	14
Max. length of posts	124
Class Distribution (%)	
Misinformation	28%
Otherwise	72%

72% non-misinformation) and a total of 30,385 posts. Each user has an average of 5 posts, with the highest number being 542. Additionally, each post averages 14 tokens, with a maximum of 124 tokens (Table 2).

4 Methodology

Problem Definition: Given a collection of posts made by the u-th user, represented by P_u, containing T pairs of text and visuals made by an X user. We aim to identify a social media (X in our case) that spreads vaaccine misinformation.

Overview of proposed architecture: Figure 2 illustrates the overall architecture of VaxMine. We present inverse operation-based cooperative multi-agent policy gradients that employ a centralized training framework with decentralized execution by applying a centralized critic and differentiated advantages. Policy gradient agents are used in text and visual selectors to determine if a feature should be selected based on the inputs. The gradient estimates from the critic are utilized to train the selectors. The differentiated advantages are reflected in rewards that compare the current global reward with those obtained when each agent's action is replaced by an opposing action. The domain-specific language model, i.e., COVID Twitter BERT (CT-BERT) [27] and vision transformer (ViT) [10] are used to obtain the text and visual features, respectively. The classifier classifies a user based on the features selected by agents.

4.1 Features Extraction

4.1.1 Textual Feature Extraction: Each text in a post is composed of a sequence of words $w_{t_1}, w_{t_2}, \cdots, w_{t_n}$, where $w_{t_i} \in R^d$ is the d-dimensional vector representing the i-th word in the t-th text, and n is the text's length. We used CT-BERT to compute the continuous representation of posts (h^{text}). CT-BERT have the same architecture as BERT but is trained on COVID-related posts.

4.1.2 Visual Feature Extraction: ViT is used to extract visual features in our method (h^{visual}). ViT applies a transformer architecture to visual patches. Position embeddings are incorporated for each fixed-size patch, and the resulting vector sequence is fed to a classic transformer encoder.

Above generated textual (h^{text}) and visual features (h^{visual}) are then fed to the inverse operation-based cooperative MARL module.

4.2 Inverse Operation Based Cooperative MARL

Motivated by [17], we introduce two agents (i.e., text selector and visual selector) that we use to select a vaccine relevant text and visual from users' historical multimodal posts.

Text Selector and Visual Selector: In one event, one user's historical posts P_u belong to the user's sequential posts. We used the features extracted h^{text} and h^{visual} for the text and visual selectors in step $t \in T$. We use $a_t \in A = \{0, 1\}$ to detect whether or not to select the feature at step $t \in T$. During implementation, the local action observation histories must be used to learn the policy $\pi(a : T)$. To understand the full history of the selectors, we used Bidirectional Gated Recurrent Unit (BiGRU) [8] to model them. As a result, the agents' policy $\pi(a : T)$ is formulated as:

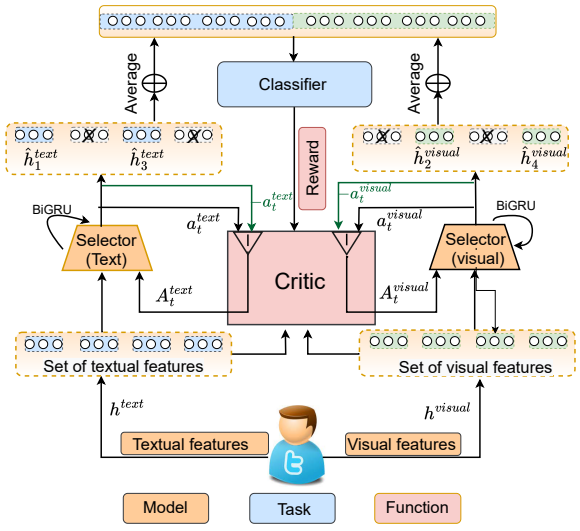

Figure 2: An overview of our proposed architecture.

$$\pi^e(a_{1:T}) = \prod_{t=1}^{T} \pi^e(a_t|s_t)$$
$$g_t^e = BiGRU(h_t^e, g_{t-1}^e) \tag{1}$$
$$\pi^e(a_t|s_t) = (1 - a_t^e) * (1 - \sigma(MLP(g_t^e))) + a_t^e * \sigma(MLP(g_t^e)),$$

$e \in \{text, visual\}$. The BiGRU's hidden state is denoted by g_t, and MLP denotes a multilayer perceptron layer. $\sigma(\cdot)$ is a sigmoid function that turns g_t into a probability. Following that, the selector samples an action to determine whether to select the feature ($a_t = 1$) or not ($a_t = 0$). The feature h_t^e will be modified as $\hat{h}_t^e$ and added to H_{indi}^e if a particular feature is selected. H_{indi}^e, a subset of the user representation, is then used to make predictions.

Classifier: For binary classification, a subset of selected features is used at each step t in one event. We combined H_{indi}^{text} and H_{indi}^{visual} using average operation to generate a user representation. Thus, we used two fully connected layers and a dropout operation to process users' representation. After the final layer's output, a non-linear sigmoid function is applied to generate the probability distribution.

$$\hat{o}_t = \text{MLP}(\text{avg}(H_{\text{text}_i}) \oplus \text{avg}(H_{\text{visual}_i}))$$
$$P_\gamma(y = \hat{y}_u|o_t; \theta_d) = \hat{y}_u\sigma(o_t) + (1 - \hat{y}_u)(1 - \sigma(o_t))$$

$\oplus$ indicates the concatenation operation and $\hat{y_u}$ denotes the probability distribution of the prediction. We can reward the selector based on the likelihood of the ground truth, i.e., $P_\gamma(y = \hat{y_u})$ to take better actions. The actor-critic technique can be used to apply the change in the $P_\gamma(y = \hat{y_u})$ after upgrading its sets with the newly selected instances as the unified temporal difference error [41].

$$\gamma_t = P_\gamma(y = \hat{y_u}|o_{t+1}) - P_\gamma(y = \hat{y_u}|o_t)$$
$$\text{L}_t(\theta_c) = [\gamma_t + (H_{indi}^{t+1}, \Pi_{t+1}, A_{t+1}) - Q(H_{indi}^t, \Pi_t, A_t)]^2 \tag{2}$$

$H^t_{indi} = H^{text}_{indi} \oplus H^{visual}_{indi}$, $\Pi = \pi^{text} \oplus \pi^{visual}$, and $A = a^{text} \oplus a^{visual}$. Whereas using similar advantages makes it hard to conclude the contribution of each selector. As a result, differentiating the advantages is essential.

Differentiated Advantages via Inverse Operation: In our proposed settings, different rewards can be implemented using a centralized critic. Although our method utilizes a centralized critic to estimate Q-values for joint actions based on the central state H^t_{indi}, we can offer each agent a unique benefit by comparing the global reward to the reward obtained when the agent performs the contrary action. Essentially, by subtracting the Q-value of the gold action from the Q-value of the opposite action, this approach generates a positive reward. Formally, we can then calculate an advantage function for each selector e by comparing the current action's Q-value a^e to an inverse operation baseline that takes the opposite action a^e while maintaining the other agent's action a^e constant:

$$A^e(H, \pi, A) = Q(H, \Pi, (A^e, a^{-e})) - Q(H, \Pi, (-a^e, a^{-e})) \quad (3)$$

As a result, $A^e(H, \Pi, A)$ calculates a baseline and a centralized critic for each agent and each advantage. Thus, Algorithm 1 can be used to optimize the model further.

Algorithm 1 Inverse Operation Based Cooperative MARL

Initialize the critic network randomly $Q(S, \pi, \mathbf{a}|\theta_Q)$ and 2 selectors $\pi(s|\theta^e_\pi)$ with weights θ_Q and θ^e_π.
Initialize target network Q' and π' with weights $\theta_{Q'} \leftarrow \theta_Q$, $\theta^e_{\pi'} \leftarrow \theta^e_\pi$. Replay buffer initialization R
for event = 1, M **do**
  Obtain the initial observation state h^e_1
  for $t = 1, T$ **do**
    Select action $a^e_t = \pi(h^e_t|\theta^e_\pi)$ as per the current policy
    Perform action a^e_t and observe the likelihood of ground truth $\mathrm{P}_Y(y = \hat{y}_u|o_t)$ and observe the new state h^e_{t+1}
    Perform action a^e_{t+1} and observe the likelihood of ground truth $\mathrm{P}_Y(y = \hat{y}_u|o_{t+1})$, as a result, obtain the reward $r_t = \mathrm{P}_Y(y = \hat{y}_u|o_{t+1}) - \mathrm{P}_Y(y = \hat{y}_u|o_t)$
    Store transition $(H^t_{indi}, A_t, r_t, H^{t+1}_{indi})$ in R
    Sample N random transitions from a minibatch $(H^i_{indi}, A_i, r_i, H^{i+1}_{indi})$ from R
    Set $z_i = r_i + \gamma Q'(H^{i+1}_{indi}, \Pi_{i+1}, A_{i+1})$
    Minimize the loss to update the critic:

$$\mathcal{L}(\theta_Q) = \frac{1}{N} \sum_i \left[z_i - Q(H^i_{indi}, \Pi_i, A_i|\theta_Q)\right]^2$$

    Updating selectors with differentiated advantages:

$$A^e(H, \Pi, A) = Q(H, \Pi, A) - (H, \Pi, (-a^e, a^{-e}))$$
$$\nabla_{\theta^e_\pi} J(\theta^e_\pi) = \nabla_{\theta^e_\pi} \log \pi(a^e_t|h^e_t) A^e(H, \Pi, A)$$

    Updating target networks:

$$\theta_{Q'} = \tau\theta_Q + (1-\tau)\theta_{Q'}, \theta^e_{\pi'} = \tau\theta^e_\pi + (1-\tau)\theta^e_{\pi'}$$

  end for
  Minimize cross-entropy loss:

$$J(\theta_C) = -\left[y_u \log \hat{y}_u + (1 - y_u) \log (1 - \hat{y}_u)\right]$$

end for

5 Experiments

5.1 Baselines

We compared our method with state-of-the-art (SOTA) unimodal (text only or visual only) and multimodal methods including recent LMMs. Our comparative methods include those used to identify vaccine misinformation and other multimodal classification tasks.

5.1.1 *Unimodal Models.*

- **Text only**: We adopted two different methods to encode textual features. (i) We concatenated all historical posts and used LSTM [19], GRU [7], a BERT [9] and GPT-4 [31] to obtain textual features. (ii) We also used DepressionNet [44], a text-only method designed to identify a user's depression using historical posts.
- **Visual only**: We used DenseNet [20], VGGNet [34], ResNet [18] and vision transformer (ViT) [10].

5.1.2 *Multimodal Models:*

- **Mid fusion methods**: For mid-fusion-based multimodal methods, we concatenated all textual and visual contents from users' historical posts and trained separate models on the textual and the visual, BERT and ResNet+BERT, respectively, and then we combined them by taking the output of the second-to-last layer of ResNet for the visual part and the output of the `[CLS]` token from BERT, and we fed them into an MLP. We also used ResNet+fasttext to extract the visual and textual features from users' historical posts and concatenated features for classification.
- **VisualBERT** [24]: We concatenated all textual and visual contents from users' historical posts and used a VisualBERT that is trained using a multimodal objective and tested on a wide range of multimodal tasks.
- **MVAE** [21]: Encoded representation of multimodal news data is fed to a MVAE to detect fake news.
- **EANN** [38]: Uses textual and visual features to train an event-based discriminator to identify fake news using social media multimodal data.
- **SDM** [3]: FastText embeddings are employed to encode historical user posts, fed into an LSTM layer following an attention layer for suicide identification.
- **UPFD** [11] detects fake news by modeling social context, and historical posts of a social media user.
- **MTAN** [6]: BERT and InceptionResNet are used to obtain features of historical textual and visual content. Extract features of historical visual and textual data are then fed to a time-aware LSTM layer, followed by attention for depression detection.
- **Large Multimodal Models (LMMs)**: For LMMs, we used LLaVA [25], MMGPT [16] and CogVLM [37].
- **DGExplain** [33]: a generative approach for detecting multimodal COVID-19 misinformation that assesses the interplay between visual and textual elements in multimodal content.
- **MASN** [15]: a tailored reinforcement learning approach to integrate the opinion-word and visual-region selection strategies to select information for opinion-word and visual-region features for multimodal personality classification.

- **Seta-Attn** [39]: We also compared results with Seta-Attn, a recently proposed method for identifying vaccine misinformation. Seta-Attn uses visual and textual content and semantic- and task-level attention to focus on the essential contents of a post that indicate vaccine misinformation multimodal posts.

5.2 Experimental settings

We reported an average of 10-fold cross-validation for all the results. We used a similar experimental setup for all baseline methods, followed the original model settings described in their paper, and used grid search optimization to obtain the optimal hyperparameters. We employed the base version of pretrained language models and utilized the Adam optimizer with a learning rate of 0.001 for model optimization. Posts of varying lengths were padded, and the model was trained for 150 epochs with early stopping configured to a patience of 10 epochs.

6 Results

Comparison with Baselines: Table 3 shows our method's performance relative to SOTA approaches. As anticipated, text-only methods outperformed visual-only methods, since text typically contains more explicit vaccine misinformation than visuals. In our experiments, we observed that using BERT for single posts results in an F1-Score of 0.60, whereas incorporating historical posts increases the F1-Score to 0.71. Additionally, transformer-based models like BERT, CT-BERT, and GPT surpassed GRU and LSTM models, which is expected due to their superior ability to capture contextual information. DepressionNet performed better than other text-only baselines, including GPT, likely due to its enhanced capability to understand users' historical posts. Furthermore, we observed that both text-only and visual-only models have relatively low performance, with F1-Scores not exceeding 73%. This suggests that relying only on text or visuals is less effective in multimodal vaccine misinformation detection. We then show how integrating both text and visuals improves performance in identifying vaccine misinformation.

We noted that multimodal methods outperformed both text-only and visual-only approaches (Table 3). Further, the SOTA attention-based multimodal methods, i.e., VisualBERT and Seta-Attn, performed slightly better than ResNet + BERT and ResNet + fasttext. However, according to our analysis, only a small percentage of users' historical posts contain relevant content to identify vaccine misinformation. As a result, multimodal methods based on attention may have difficulty capturing the relevant content. We adopted an RL-based selection strategy to address the above challenge. The proposed method reaches an F1-Score of 0.94, marking a 13% absolute improvement over Seta-Attn (the top baseline), which is developed for detecting vaccine misinformation using multimodal data. Our results validate that differentiated advantages outperform attention-based (Seta-Attn) multimodal methods. We attribute the performance gains to selecting relevant content from users' historical posts, and the use of differentiated content also improves performance compared to attention-based methods for identifying vaccine misinformation. We also note that the effectiveness of other multimodal methods, including LMMs (i.e., MVAE, EANN, SDM, UPFD, DGExplan, MTAN, CogVLM, LLaVa and MMGPT) are less desirable in detecting vaccine misinformation on multimodal social media posts. The reason is that these methods encode both relevant and irrelevant content from the users' historical multimodal posts on social media that add noise and result in low performance.

Table 3: Performance comparison. * indicates that our method achieved a significant ($p < 0.05$) improvement than the best baseline (underlined) according to the Mann–Whitney U test.

Modality	Method	F1-Score	Precision	Recall
Text	GRU	0.69	0.68	0.70
	LSTM	0.68	0.69	0.66
	BERT	0.71	0.70	0.71
	CT-BERT	0.72	0.71	0.71
	GPT-4	0.72	0.70	0.72
	DepressionNet	0.73	0.72	0.73
Visual	DenseNet	0.65	0.71	0.64
	ResNet	0.66	0.68	0.66
	VGGNet	0.64	0.72	0.65
	ViT	0.67	0.74	0.67
Multimodal	ResNet+BERT	0.75	0.75	0.76
	ResNet+fasttext	0.74	0.78	0.71
	VisualBERT	0.77	0.76	0.72
	MVAE	0.78	0.77	0.78
	EANN	0.79	0.78	0.79
	SDM	0.77	0.77	0.77
	MTAN	0.79	0.78	0.79
	CogVLM	0.78	0.78	0.78
	LLaVA	0.80	0.78	0.79
	MMGPT	0.77	0.77	0.78
	UPFD	0.79	0.79	0.79
	DGExplain	0.80	0.79	0.80
	MASN	0.78	0.80	0.79
	Seta-Attn	0.81	0.81	0.81
	Proposed	0.94*	0.94*	0.94*

Post-wise comparison: Figure 3 illustrates the performance of our method with different numbers of posts compared to the top baseline method (Seta-Attn). Our method consistently increases F1-Score and consistently surpasses Seta-Attn across all evaluated post quantities. We also found that our method achieves the highest F1-Score with the first 5 posts, representing the average number of historical posts per user in our dataset. In contrast, Seta-Attn's F1-Score rose from 62% to 80% with the first 20 posts and reached a peak of 81% with all historical posts. Our method, however, maintains stable performance, and increasing the number of historical posts does not significantly improve the model's performance. The observed performance saturation is attributed to the small number of users with more than 20 posts.

6.1 Analysis

Ablation analysis: An ablation analysis (Table 4) demonstrates how each new component we incorporated into our method enhanced the overall performance. The F1-Score drops (from 0.94 to 0.86) when we remove the selectors and feed both relevant and irrelevant multimodal content, indicating the importance of selecting

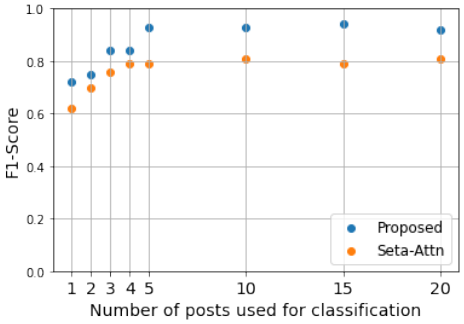

Figure 3: Post-wise analysis: Proposed v/s Seta-Attn

texts and visuals that are more useful to classify vaccine critical users. This drop in F1-Score validates our motivation to select relevant content, which is ignored in previous studies. Further, we first removed visual feature extraction from our method to demonstrate the importance of multimodal features. The performance dropped to 0.76 when we removed the text feature module from our proposed method. These results show that textual information plays a more crucial role than visual information in identifying vaccine critical users. Similarly, the F1-Score drops to 0.84 when removing the visual feature extraction module from our proposed method. This drop in performance shows the importance of using text and visual content and validates our motivation that both visual and textual content should be jointly considered to make accurate inferences. Therefore, we conclude that the advantages of the proposed method are due to the integration of multimodal features and the selection of relevant posts from a user's history, which together enhance performance.

Table 4: Ablation analysis: "Proposed w/o selector" demonstrates the result of using all posts, meaning without the selector module. "Proposed w/o visual features" and "Proposed w/o text features" display the results when visual and text features are excluded from our method. *indicates that our method achieved a significant ($p < 0.05$) improvement over other variants of our method according to the Mann–Whitney U test.

Method	F1-Score	Precision	Recall
Proposed	0.94*	0.94*	0.94*
Proposed w/o selector	0.86	0.86	0.86
Proposed w/o visual features	0.84	0.84	0.85
Proposed w/o text features	0.76	0.81	0.78

Robustness analysis in realistic settings: Considering realistic settings where there could be a small percentage of vaccine critical users, we conducted a robustness analysis of the proposed method on different percentages of vaccine critical users (Figure 4). To conduct the robustness analysis, we changed the ratio of users from 10% to 90% at increments of 10% and observed that our method achieved better performance despite of low percentage of vaccine critical users. We also note that when the percentages do not reach 50%, the best baselines method (Seta-Attn) performs poorly. Our

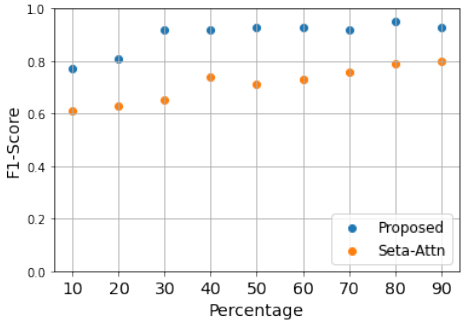

Figure 4: Proposed v/s Seta-Attn trained on data with varying percentages of vaccine critical users to validate the robustness.

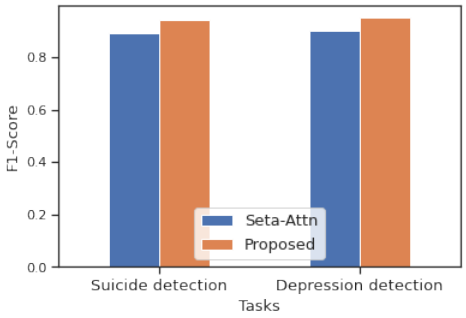

Figure 5: Generalizability: proposed v/s the baseline method on Suicide detection and Depression detection task.

method outperformed Seta-Attn with an F1-Score of 77% (an absolute increase of 16%) when tested on 10% of the data. Hence, we can conclude that our method is robust in realistic scenarios.

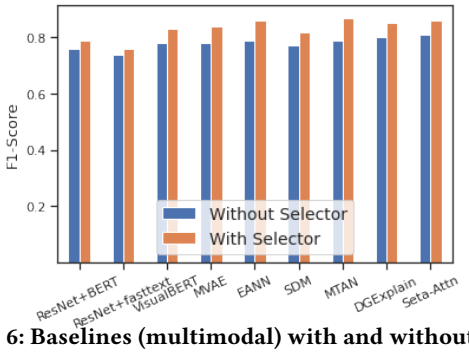

Figure 6: Baselines (multimodal) with and without selector

Generalizability test: A generalizability test on suicide detection [3] and depression detection [6] tasks[2] that use multimodal data with the user's historical posts shows that our method outperformed the best results reported in [3] for suicide detection and in [6] for depression detection and Seta-Attn by an absolute increase of 4% and 2% on both tasks (Figure 5). Our generalizability test concludes that our method is generalizable and performs better than the best baseline in suicide and depression classification tasks.

Multimodal baselines with selector: We also investigated the effectiveness of our selector module in multimodal baselines (Figure 6). We noticed an improvement in the performance of each

[2]We used these due to the unavailability of vaccine misinformation multimodal datasets

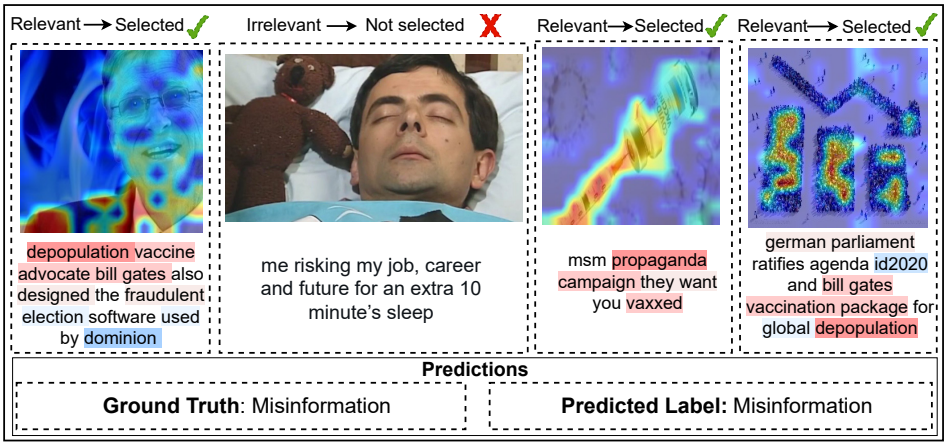

Figure 7: Relevant posts are selected and visualized using Gradcam and SHapley Additive exPlanations (SHAP) [26], where high visual scores (important words) are represented by red, whereas low visual scores (less important words) are in blue. The bottom row shows the prediction results using our method.

tested baseline, with increases ranging from 2% to 8%. Seta-Attn obtained the highest F1-Score of 0.86 after incorporating a selector module that selects the relevant content from the user's historical posts. This improvement in baseline results confirms the usefulness of using a selector that selects relevant content from the user's historical posts.

Qualitative analysis: Figure 7 shows an example correctly predicted by our method because it selects vaccine relevant posts from users' historical multimodal posts and focuses on the important features. For example, for visuals, our method captures the face (of Bill Gates), a syringe, and clusters of people (showing the decrease in population). For text, words highlighted in red such as depopulation, vaccine advocate, Bill Gates, propaganda, etc., contribute more to the final prediction. We also show that selector module selects relevant content and ignores the irrelevant content (user post #2) from the user's historical posts. Further, from the prediction analysis, we observed that our method correctly classified a user by selecting relevant posts and leveraging visual and textual data.

Error analysis: When we examined specific examples of incorrect predictions, we found they mostly corresponded to one of three types. These were posts with insufficient information, such as when neither the visual nor the text contained anti-vaccine words or opinions, when OCR failed to detect any words, and when the posts required external knowledge.

Practical Implication: Social media posts that include text and visuals can spread quickly and may contribute to normalizing vaccine misinformation, as well as making them seem more common than they are. Tools for quickly determining which users are posting vaccine misinformation at scale are important for catching the spread of vaccine critical content, including misinformation and other efforts at undermining public health actions. Tools that make use of methods for classifying users and posts from multimodal data can then support public health organizations through early signaling and identification of emerging misinformation threats, which in turn can guide the design and deployment of countermeasures delivered via social media.

Our results show that we can accurately predict which users are engaging with and spreading vaccine misinformation by looking at their historical posts. This can support actions taken by social media platforms, including precise flagging of posts so that other social media users are warned about the content in advance or reducing the visibility of the content by downranking or not recommending it in user timelines. We expect that this work will have an impact beyond vaccination and could be valuable in other situations where multimodal data is used to disseminate misinformation.

7 Conclusion

We investigate the challenge of identifying vaccine misinformation on X. Our contribution is to release a new multimodal dataset (MM-Vax) with historical posts of 2,072 X users and a novel reinforcement learning-based method (VaxMine) that selects the relevant content from the historical posts of a user for better classification performance. Our experimental results showed that our method outperformed SOTA methods. We further showed the generalizability of our method on other multimodal tasks. We demonstrated that understanding both modalities and selecting relevant content from users' historical posts is crucial to detecting vaccine misinformation on social media.

Ethical Considerations

We carefully addressed potential ethical issues in this study, focusing on (i) protecting user privacy and (ii) avoiding harmful uses of the proposed dataset. X's privacy policy explicitly allows third parties to copy user content through the X API. We follow established social media research ethics guidelines, which permit the use of user data without explicit consent as long as anonymity is preserved [1, 40]. We did not collect any metadata that could identify authors. Given the subjective nature of annotation, some biases in the distribution of labels and our gold-labelled data are expected; however, these biases are unintentional. Additionally, all content is manually reviewed to remove personally identifiable information. The released annotated data includes de-identified, publicly accessible posts from X, where users are aware that their content is publicly available and have no expectation of privacy. Thus, no ethical approval is needed.

Acknowledgments

Matloob Khushi is supported by UKRI NERC grant NE/X000192/12; Jinman Kim by Telehealth and Technology Centre, Nepean Hospital.

References

[1] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 94–102.

[2] Taxiarchis Botsis, Michael D Nguyen, Emily Jane Woo, Marianthi Markatou, and Robert Ball. 2011. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association* 18, 5 (2011), 631–638.

[3] Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1718–1728.

[4] 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*.

[5] 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.

[6] Ju Chun Cheng and Arbee LP Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems* (2022), 1–21.

[7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

[8] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[9] 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

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2051–2055.

[12] Jingcheng Du, Jun Xu, Hsingyi Song, Xiangyu Liu, and Cui Tao. 2017. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of biomedical semantics* 8, 1 (2017), 1–7.

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

[14] 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.

[15] Xiaoya Gao, Jingjing Wang, Shoushan Li, Min Zhang, and Guodong Zhou. 2022. Cognition-driven multimodal personality classification. *Science China Information Sciences* 65, 10 (2022), 202104.

[16] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).

[17] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in Twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 110.

[18] 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.

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

[20] 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.

[21] 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.

[22] Ray R. Larson. 2010. Introduction to Information Retrieval. *J. Am. Soc. Inf. Sci. Technol.* 61, 4 (April 2010), 852–853. https://doi.org/10.1002/asi.v61:4

[23] Stephan Lewandowsky, John Cook, Philipp Schmid, Dawn Liu Holford, Adam Finn, Julie Leask, Angus Thomson, Doug Lombardi, Ahmed K Al-Rawi, Michelle A Amazeen, et al. 2021. The COVID-19 Vaccine Communication Handbook. A practical guide for improving vaccine communication and fighting.

[24] 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).

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[27] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020).

[28] 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

[29] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive GRU. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[30] 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.

[31] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2, 5 (2023).

[32] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021. Towards Ordinal Suicide Ideation Detection on Social Media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 22–30.

[33] 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.

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

[35] APOORVA SINGH, Soumyodeep Dey, Anamitra Singha, and Sriparna Saha. 2022. Sentiment and Emotion-Aware Multi-Modal Complaint Identification. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12163–12171. https://doi.org/10.1609/aaai.v36i11.21476

[36] 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.

[37] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023).

[38] 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.

[39] 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.

[40] Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 51, 6 (2017).

[41] Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, and Li Fei-Fei. 2017. Learning to learn from noisy web videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5154–5162.

[42] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal Vlog Dataset for Depression Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12226–12234. https://doi.org/10.1609/aaai.v36i11.21483

[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] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: learning multi-modalities with user post summarization for depression detection on social media. In *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 133–142.