GatedVAE: Detecting Multimodal Fake News with Gated Variational AutoEncoder

Yimeng Gu
Queen Mary University of London
United Kingdom
yimeng.gu@qmul.ac.uk

Ignacio Castro
Queen Mary University of London
United Kingdom
i.castro@qmul.ac.uk

Gareth Tyson*
The Hong Kong University of Science and Technology (GZ)
China
gtyson@ust.hk

ABSTRACT

This paper focuses on the challenge of automatically detecting *multimodal* fake news on social media. Although multimodal fake news classifiers exist, we show that prior works fail to reflect certain real-world practicalities. For example, news captions often contain highly irrelevant information that introduces noise to the overall message contained within the post. Existing classifiers do not properly address this, resulting in misclassifications. To address this limitation and suppress noise, we propose GatedVAE (Gated Variational AutoEncoder), which enables VAE with the gating mechanism. Experimental results demonstrate the efficacy of our approach: GatedVAE is able to suppress noise and learn an effective multimodal representation. It outperforms state-of-the-art models by 3.7% and 2.4% (F1) on Twitter and Weibo datasets, respectively. Our ablation study highlights the importance of the gating mechanism and the methods we adopt to alleviate overfitting. We further show that, in addition to dynamically controlling the pass of noisy input, the gate also helps to uncover modality importance in multimodal fake news detection.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Social networks**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

fake news detection, multi-modal learning, adaptive gated mechanism

ACM Reference Format:
Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2024. GatedVAE: Detecting Multimodal Fake News with Gated Variational AutoEncoder. In *ACM Web Science Conference (Websci '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3614419.3643992

1 INTRODUCTION

It is now common for users to obtain news from social media sites, such as Twitter, Facebook and Youtube [2]. Whilst in 2016, 62% of adults in the US got their news from social media, this figure has since increased to 71% in 2021 [1]. Such trends have been accompanied by an equivalent expansion in the volume of so-called "fake news" circulating on social media. The nature of this material is diverse, ranging from whimsical (*e.g.* The Onion) to life threatening (*e.g.* anti-vax hoaxes) [36, 45]. Consequently, it has gained significant attention, with social media companies actively attempting to build models that can detect such content.

Building accurate models is not trivial though. A key challenge in detecting fake news is the growing diversity of such material. For example, news posted on social media covers a vast array of topics and often consists of diverse modalities (*e.g.* text, image and video). Consequently, it is increasingly important to design classifiers that can leverage *both* language and visual modalities. A small set of prior studies have considered this challenge [18, 35, 38]. However, to date, they have taken relatively simple approaches, *e.g.* encoding language and visual inputs [18, 35, 38], and then applying fusion methods (e.g. concatenation [38], cross-modal attention [9]). Yet, these techniques treat both modalities as equally important, failing to suppress noise that may emerge within the individual modalities. For example, news posts often contain highly emotive (but irrelevant) words that introduce significant noise to the overall message contained within the post (see Tab. 1 for examples). In this paper, we define *noise* as non-useful information for determining the credibility of news, opposite to *signal* which contains crucial information that can be exploited for classification.

Our experiments show that this noise poses a significant challenge for simple feature representations. In our later experiments (Sec. 3) we separately and combinatively perturb the text and image of news posts on Twitter to monitor the impact it has on several state-of-the-art (SOTA) multimodal fake news detectors. We find that the impact of textual perturbation on the classification results is negligible — even when the perturbations change the entire meaning of the news, the model outputs are invariant. This is particularly worryingly as it allows adversaries to easily bypass such detectors. Thus, we argue that it is vital to explore new ways for multimodal fake news classification to filter out these noisy modality inputs, whilst retaining useful information.

Our paper has three key objectives: (*i*) to devise an effective way of representing multimodal data, particularly oriented towards fake news detection; (*ii*) to adaptively suppress the noisy feature inputs that contribute less to the news label prediction (fake or real); and (*iii*) to build a classifier that can exploit these adapted feature inputs to detect fake news. To achieve our objectives, we propose GatedVAE, a noise-aware multimodal fake news classifier. GatedVAE adopts a variational autoencoder (VAE) as the multimodal feature

*Also with Queen Mary University of London.

Websci '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 979-8-4007-0334-8/24/05…$15.00
https://doi.org/10.1145/3614419.3643992

Table 1: Examples of noisy news text in the Twitter dataset

News Text	News Event	Type of Noise
This shit crazy ‼ #newyork #subway #flooded #hurricanesandy #timessquare	Sandy Hurricane	Sentimental words
LMFAO‼! [laugh-cry-emoji] #Sandy	Sandy Hurricane	Emojis and abbreviations
http://t.co/pS4KetWGaW #4chan	Boston Marathon	No substantial text

extractor. We show that it can effectively learn latent vectors, representing both modalities in tandem. Our results, however, show that this alone is ineffective at suppressing noise. Thus, we further incorporate a gated mechanism, to adaptively block out noisy inputs. Our evaluation shows that our approach is able to decide which modality is "useful", such that the noisy inputs can be blocked. GatedVAE outperforms SOTA models by 3.7% and 2.4% (F1) on Twitter and Weibo datasets, respectively. Importantly, it outperforms SOTA by 10.6% in recall (on Twitter dataset [6]). Note, recall is arguably more important for fake news detection as it is possible for human moderators to review any incorrect classifications.

In summary, the contributions of this paper are as follows:

(1) We explore the deficiencies of existing multimodal fake news models through a set of perturbation experiments (Sec. 3). We show that all 3 tested models are insensitive (avg. Δ Acc = 1.26%) to text perturbations, yet sensitive (avg. Δ Acc = 10.36%) to image perturbations. We conjecture that one possible explanation is that there is significant noise (Tab. 1) in textual inputs, leading existing models to "ignore" textual modality as a whole and become ineffective at differentiating useful signals from noise.
(2) Exploiting this observation, we propose GatedVAE, a gated variational autoencoder (Sec. 4) that suppresses noisy unimodal inputs from multimodal fake news. We explore three types of gated module and explore their efficacy: unimodal soft gate, multimodal soft gate and reinforcement learning (RL)-enabled on/off gate.
(3) We evaluate GatedVAE (Sec. 6) on two benchmark datasets from Twitter [6] and Weibo [16] (Sec. 5). GatedVAE outperforms the SOTA models by 3.7% and 2.4% (F1) on Twitter and Weibo dataset, respectively. In addition, it outperforms SOTA by 10.6% in recall on a Twitter dataset.
(4) We show that the proposed gates also help to uncover the more important modality when detecting a multimodal fake news item. Further, we evaluate and compare the efficacy of the gates by conducting an ablation study (Sec. 6).

2 RELATED WORK

There have been a number of recent works that attempt to automate fake news detection, including multimodal techniques. Here, we summarize existing fake news detection approaches and gated fusion methods.

2.1 Fake News Detection

Prior efforts [3, 7, 40] have adopted traditional machine learning classifiers, e.g. SVMs, to detect fake news. Later works focus on capturing semantic information of the news using recurrent neural networks (RNNs) [25] and convolutional neural networks (CNNs) [44], further coupled with attention mechanisms [13, 29]. Other works take advantage of pre-trained language models [17, 28] followed by simple classifiers (achieving promising performances). In addition to detecting fake news based on its semantic meaning, some works attempt to differentiate fake news and real news using the news propagation network [5, 14] or comment-reply network [42]. These approaches are effective in detecting fake news, but they only based on unique modality and cannot accommodate multimodal inputs.

2.2 Multimodal Fake News Detection

To detect fake news in a multimodal setting, it is important to first find effective ways to encode news' textual and visual inputs. Many techniques have been used to learn unimodal representations: [18] uses multimodal variational auto-encoder; [22] uses CNNs and RNNs with attention; [35] leverages pre-trained models, followed by shallow neural layers. With the learned unimodal representations in hand, the next step is to form the multimodal representation. Adopted fusion methods include feature concatenation [9, 18, 20, 27, 35, 38], element-wise multiplication [22] and cross-modal attention [16]. Apart from effectively representing news inputs, many works incorporate co-learning task(s) tailored to improve performance on fake news detection. [38] relies on adversarial networks. [9] measures cross-modal ambiguity to adaptively aggregates unimodal features and cross-modal correlations. [27] trained a reinforcement learning agent to enable domain adaptive fake news detection.

The above works obtain promising results in multimodal fake news detection, but assume that both modalities contribute equally to the news label prediction, neglecting the fact that text posted on social media often contain unrelated information. Although [34] uses a multiplicative multimodal method to decay the decision signal coming from the noisy modality, they delay the signal suppression at the decision point. In contrast, our work is able to suppress noisy modality *inputs* at an earlier stage, which helps to rapidly pinpoint to the problematic modality that may cause the news to be fake. In summary, there is little work taking into account the noise contained in input modalities. In our work, we strive to build a solution that can dynamically suppress the input from a noisy modality. To enable this, we take the advantage of gated fusion.

2.3 Gated Fusion

Prior works [31, 46] have shown that a gated module can be used to understand feature importance for multimodal fusion. Many works implement the soft gate, whose output is a continuous numerical value. This has been used to assist various computer vision tasks, *e.g.* image dehazing [46] and saliency prediction [21]. Moreover, [37] proposes a cross-gating strategy using linear layers in video

captioning task. Likewise, [11] uses the gate to avoid introducing useless inputs from images into multimodal sentiment classification. [24] presented group gated fusion to learn the contribution of each representation. Other works implement the on/off gate, which outputs either 0 or 1 as the multiplier. For instance, [8] proposes a gated controller learned through reinforcement learning to dynamically decide whether to block audio and video modality at each timestamp. Inspired by [8], we adopt and modify their on/off gate in order to applying it to our non-temporal multimodal inputs. Our work is the first to introduce a gated module to suppress noisy modality input in multimodal fake news detection.

3 MOTIVATION

Given that news posted on social media tends to be noisy (containing unrelated information, see Tab. 1) for classification, we hypothesize that it may lead existing multimodal fake news detector to largely ignore the noisy modality's contribution. To test our hypothesis, we perform experiments on three SOTA models: EANN [38], MVAE [18] and SpotFake [35]. We run these models using their publicly available code. Note, as some code snippets are missing, our obtained results are slightly different from their original results, although we tried our best to faithfully rewrite the missing code. We conduct pilot experiments on Twitter dataset [6].

We first train each model with their default configurations on the training set. The model with optimal performance on the original testing set is retained. After that, we perturb the news' textual and visual parts in the original test set. We then evaluate the trained models' performance change on the perturbed test set.

We use seven text and image perturbation methods: **P1. Character Perturbation:** perturbs 10% of characters in each post's text by substituting, deleting, inserting, and swapping adjacent characters (using the `textattack` [26] package). Note, we also experiment with different ratios (10%, 20% and 50%) to discover the same results. Thus, for simplicity, we only include 10% here. **P2. Text Replacement:** replaces each post's text with one non-factual sentence (we repeat P2 with 5 different fake news text from the Twitter dataset, and report the averaged experimental result: *"The incredibly rare Black Lion..."*, *"David Bowie and Tilda Swinton Dressed As Each Other"*, *"Unbelievable Shot Of The 2012 Supermoon In Rio de Janeiro"*, *"Germans march, yell 'Germany is w/ France'"*, *"Lenticular clouds over Mount Fuji, Japan"*). **P3. Text Removal:** replaces each post's text with one non-meaningful word: *"blank"*, mimicking an empty sentence. **P4. Pixel-wise Perturbation:** injects adversarial perturbation (using [32]) into each post's image. **P5. Image Replacement:** replaces each post's image with one fake news image (we repeat P5 with 5 different fake news images from the Twitter dataset, and report the averaged experimental result: *black_lion_1*, *rio_moon_1*, *five_headed_snake_5*, *half_everything_7*, *bowie_david_2*). **P6. Image Removal:** replaces each post's image with a blank image. **P7. Text and Image Perturbation:** combines **P1** and **P4**.

The perturbation becomes stronger from **P1(P4)** to **P3(P6)**. Further, **P2**, **P3**, **P5** and **P6** turn all items into fake. Note, we do not change the news label when performing the perturbation. As such, **P2**, **P3**, **P5** and **P6** *should* cause a significant change in performance if the model is able to effectively identify fake news, because the preserved label will no longer match the perturbed data.

Tab. 2 summarizes the results. In line with our hypothesis, text perturbations (**P1-P3**) do *not* significantly change the model performances. For EANN, the most significant change is caused by **P2**: however, it actually causes an *increase* (not decrease) in its accuracy (of 3.7%), with a $0.9\% - 12.3\%$ increase in the other metrics. Since EANN's performance is far worse than the other two models (and unstable while we reproduce it), we believe that the 3.7% increase cannot be interpreted as a performance enhancement. For MVAE and SpotFake, the performance change in all experiments are trivial, within 1% across all metrics. Manual check reveals that the model's predicted probability score for each class remains unchanged. This is clearly problematic. For example, **P2** changes the text to a demonstrably fake news item, yet the classifiers fail to detect this fact. This suggests that the models do not take the semantic meaning of the textual part into consideration, which we did not expect.

In contrast, image perturbations (**P4-P6**) noticeably decrease the models' performances (compared to text perturbations). **P4** sees $11.3\%, 2.2\%, 3.3\%$ accuracy drops across all three models, respectively. **P5** and **P6** further downgrade the models to a classifier that only outputs *fake* as the predicted label - the precision, recall and F1 score are therefore reduced to 0 for real news. This further explains why recall for fake news surges. When injecting image and text perturbation at the same time (**P7**), the performance drop is almost the same as **P4** for all three models. Together with the textual perturbation results, this means that tested models mainly rely on visual modality for the news classification.

In summary, the results indicate that the models' outputs are insensitive to changes made in the text input. This occurs, *even when our perturbation flips the groundtruth label.* This implies that SOTA multimodal fake news detectors do not properly consider the textual input, even though some text is relevant and essential. In practical scenarios, we consider it imperative that classifiers better balance the roles of two modalities. To address this, we next explore the use of a gated module, which adaptively suppresses inputs that may contain noise for the classification task.

4 OUR PROPOSED APPROACH

In this section, we describe our proposed approach for multimodal fake news detection: GatedVAE. We first introduce GatedVAE , before discussing the training details. The overall architecture of GatedVAE is depicted in Fig. 1.

4.1 Multimodal Feature Representation

Since the purpose of a VAE is to find the latent variable that captures meaningful factors of variation in the input data [12, 23], we believe that the learned latent variable is essential for fake news detection. We leverage a multimodal VAE similar to [18] to extract the latent multimodal representation from the news. In contrast to their approach, we add modules to alleviate overfitting and add the gate to dynamically adjust the passing of modality noise. Our multimodal VAE consists of two encoders and two decoders for textual input and visual input respectively. Concretely, the encoders are used to learn meaningful unimodal representations, which are then concatenated to form a multimodal representation. After that,

Table 2: Results of the perturbation experiments on the Twitter dataset. Numbers are reported in accuracy (Acc), Precision (Prec), Recall (Rec) and F1. Δ{metric} is computed by subtracting the corresponding performance under no perturbation. →0 indicates that the performance is reduced to zero after the perturbation. ≡ indicates that the performance has not changed. NA indicates not applicable. F1 being NA happens when the Prec and/or Rec is reduced to 0.

Model	Class Label	No Perturbation				Character Perturbation				Text Replacement				Text Removal			
		Acc	Prec	Rec	F1	ΔAcc	ΔPrec	ΔRec	ΔF1	ΔAcc	ΔPrec	ΔRec	ΔF1	ΔAcc	ΔPrec	ΔRec	ΔF1
EANN	real	0.578	0.544	0.873	0.668	↑0.036	↑0.018	↑0.050	↑0.030	↑0.037	↑0.021	↑0.069	↑0.038	↑0.031	↑0.019	↓0.042	↑0.003
	fake		0.727	0.312	0.434		↑0.084	↑0.020	↑0.037		↑0.123	↑0.009	↑0.032		↓0.002	↑0.094	↑0.095
MVAE	real	0.679	0.607	0.717	0.658	↓0.001	↓0.002	↓0.001	↓0.002	≡	≡	≡	≡	≡	≡	≡	≡
	fake		0.752	0.649	0.697		≡	≡	≡		≡	↑0.001	↑0.001		≡	↑0.001	↑0.001
SpotFake	real	0.760	0.939	0.475	0.631	↑0.003	↑0.001	↑0.004	↑0.004	↑0.004	↑0.001	↑0.007	↑0.007	↑0.004	↑0.001	↑0.007	↑0.007
	fake		0.709	0.976	0.822		↑0.005	↑0.001	↑0.003		↑0.005	↑0.001	↑0.003		↑0.005	↑0.001	↑0.003
Model	Class Label	Pixel-wise Perturbation				Image Replacement				Image Removal				Text & Image Perturbation			
		ΔAcc	ΔPrec	ΔRec	ΔF1	ΔAcc	ΔPrec	ΔRec	ΔF1	ΔAcc	ΔPrec	ΔRec	ΔF1	ΔAcc	ΔPrec	ΔRec	ΔF1
EANN	real	↓0.113	↓0.087	↓0.411	↓0.208	↓0.028	→0	→0	NA	↓0.096	↓0.092	↓0.210	↓0.131	↓0.128	↓0.088	↓0.327	↓0.171
	fake		↓0.243	↑0.161	↑0.044		↓0.177	↑0.688	↑0.276		↓0.171	↑0.029	↓0.011		↓0.274	↑0.050	↓0.032
MVAE	real	↓0.022	↓0.018	↓0.016	↓0.018	↓0.127	→0	→0	NA	↓0.127	→0	→0	NA	↓0.031	↓0.033	↓0.033	↓0.044
	fake		↓0.014	↓0.016	↓0.016		↓0.200	↑0.351	↑0.014		↓0.200	↑0.351	↑0.014		↓0.030	↓0.031	↓0.031
SpotFake	real	↓0.033	↓0.042	↓0.040	↓0.042	↓0.193	→0	→0	NA	↓0.193	→0	→0	NA	↓0.045	↓0.043	↓0.044	↓0.049
	fake		↓0.030	↓0.024	↓0.029		↓0.142	↑0.024	↓0.098		↓0.142	↑0.024	↓0.098		↓0.039	↓0.033	↓0.039

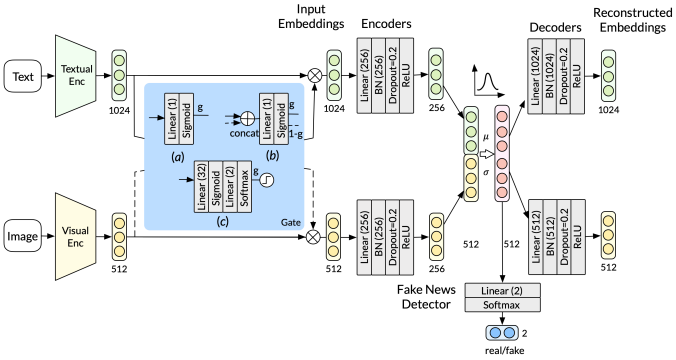


Figure 1: The architecture of GatedVAE , composed of two sets of VAE encoder and decoder, a gated module and a fake news classifier. Three types of gate modules that we implement: (a) unimodal soft gate, (b) multimodal soft gate and (c) RL-enabled on/off gate. Only implementation (b) controls both textual and visual inputs.

decoders directly reconstruct unimodal inputs from the sampled multimodal representation, to ensure that the learned multimodal representation is substantial and accurate.

4.1.1 Input Embeddings. Input embeddings generated from large pre-trained models are able to encode abundant contextual information. Furthermore, pre-trained models trained on text-image pairs are a powerful tool to capture the interconnection between visual and textual inputs. Considering the above merits, we use CLIP [30] to obtain news image embedding $\mathbf{x}_I$, and use LASER [4] to obtain news text embedding $\mathbf{x}_T$. These unimodal embeddings serve as the inputs to the encoders of our multimodal VAE (which aims to be reconstructed in decoders later on).

4.1.2 Multimodal VAE. Since the embedding already contains abundant semantic information, there is little need to further extract contextual information with complex and deep layers in VAE's encoders. Thus, we decide to use one linear layer followed by batch normalization (BN) and dropout ($p = 0.2$) as the multimodal VAE's encoder structure. BN is used to reduce internal covariance shift [15] by projecting the input to mean of zero and the variance of 1. We apply BN and dropout as they help overcome overfitting. Formally, the encoder is represented as:

$$\mathbf{z} = \mathbf{x}A^T + \mathbf{b} \tag{1}$$

$$\mathbf{z} = \frac{\mathbf{z} - E[\mathbf{z}]}{\sqrt{Var[\mathbf{z}] + \epsilon}} * \gamma + \boldsymbol{\beta} \tag{2}$$

where $\mathbf{x} \in \{\mathbf{x}_T, \mathbf{x}_I\}$, and $\mathbf{z}$ is the learned unimodal latent variable. The encoder is denoted as $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\boldsymbol{\phi}$ denotes parameters of the encoder, learned through the model training process. The encoder can be regarded as making approximate posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ close to prior $p_{\boldsymbol{\theta}}(\mathbf{z})$[23].

Now that we have the textual latent variable $\mathbf{z}_T$ and visual latent variable $\mathbf{z}_I$. The next step is to find an effective multimodal representation. In our approach, the multimodal representation $\mathbf{z}_{I+T}$ is formed through concatenating two unimodal latent variables.

In the subsequent decoding part, $\hat{\mathbf{z}}_{I+T}$ is drawn from the distribution of $\mathbf{z}_{I+T}$, subject to Gaussian distribution. It is then input to the image and text decoder respectively to reconstruct the input embeddings, $\mathbf{x}_T$ and $\mathbf{x}_I$. Symmetric to the encoder, the decoder

also consists of a linear layer followed by batch normalization and dropout. The decoder is denoted as $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, where $\boldsymbol{\theta}$ refers to parameters of the decoder. In other words, the decoder tries to generate $\mathbf{x}$ from latent representation $\mathbf{z}$.

4.2 Gated Mechanism

As mentioned earlier, the textual modality of social media can be very *noisy*, sometimes containing non-useful information. A limitation of the multimodal VAE is that it is insensitive to text perturbations made in news text. As such, due to the amount of noisy textual inputs, news text containing useful information might also be omitted by the model.

This leads us to apply a gated module to the **input embedding** of the multimodal VAE, so that it can selectively pass through meaningful textual input and suppress noisy textual input at an earlier stage. To achieve this, we implement and compare three types of gated module: (*a*) unimodal soft gate; (*b*) multimodal soft gate; and *(c)* RL-enabled on/off gate. The unimodal gate and RL gate only control the pass of the textual input, while the multimodal gate controls both textual and visual inputs. The reason why we do not implement a single gate for the visual input is that the visual modality is less noisy (Sec. 3). We explore the above three gates in order to comprehensively understanding the gate design space and select one that best suits our application scenario. We next describe these three approaches in detail.

4.2.1 Unimodal Soft Gate. The unimodal soft gate is only applied to the textual inputs. It is composed of a linear layer and a sigmoid function (shown in Eq.3).

$$\sigma(x) = \frac{1}{1 + exp(-x)} \tag{3}$$

Given the input $\mathbf{x}_T$, the gate $G(\mathbf{x}_T; \boldsymbol{\theta})$ computes a score g between 0 and 1, which is then multiplied to the original input $\mathbf{x}_T$ in Eq. 4. If the gate output is 0, the text input is totally blocked; if the gate output is 1, the text input is totally passed through. Any value in the middle can be interpreted as a weight assigned to the text input.

$$\mathbf{x}'_T = g * \mathbf{x}_T = G(\mathbf{x}_T; \boldsymbol{\theta}) * \mathbf{x}_T \tag{4}$$

4.2.2 Multimodal Soft Gate. Similar to the unimodal gate, our multimodal soft gate also consists of a linear layer and a sigmoid function. However, here the input is the concatenated textual and image embeddings. Furthermore, both the textual input and image input is controlled by the gate. Formally, the gate can be expressed as:

$$\begin{aligned} \mathbf{x}'_T &= g * \mathbf{x}_T = G([\mathbf{x}_T; \mathbf{x}_I]; \boldsymbol{\theta}) * \mathbf{x}_T \\ \mathbf{x}'_I &= (1 - g) * \mathbf{x}_I = (1 - G([\mathbf{x}_T; \mathbf{x}_I]; \boldsymbol{\theta})) * \mathbf{x}_I \end{aligned} \tag{5}$$

4.2.3 RL-enabled on/off Gate. Finally, we implement an RL-enabled on/off gate to simulate the hard pass/stop of text inputs. The gate consists of two linear layers and softmax function. If the gate outputs 1, then the text signal is unaffected. Otherwise the gate outputs 0, which means the text signal is entirely replaced by a zero vector.

$$\mathbf{x}'_T = \begin{cases} \mathbf{x}_T & \text{if} \quad g = 1 \\ \mathbf{0} & \text{if} \quad g = 0 \end{cases} \tag{6}$$

Algorithm 1 Training on/off gate

```
for epoch ← 1 : epochs do
    for k ← 1 : n do
        p_pass ← predict(g, x_T)
        x'_T ← 0
        x'_T ← x_T with probability p_pass
        loss_k ← trainVAE(x'_T, x_I)
    end for
    updateGate(gate, loss_k, loss_baseline)
    updateLossBaseline(loss_k, loss_baseline)
end for
```

The soft gate can be trained through gradient descent methods because it has a continuous mathematical formulation. However, we cannot use gradient descent methods to update the on/off gate's parameters. This is because the set of possible values of g is discrete. Instead, we use the policy gradient descent method to train the on/off gate. Following [8], we take $e^{-\mathcal{L}}$ (where $\mathcal{L}$ is the cross entropy loss (described in Sec. 4.3) from news label prediction) as the reward signal, since smaller training loss indicates better model performance. We are maximizing the expected reward as represented in Eq. 7:

$$J(\boldsymbol{\theta}_g) = E_{P(g|\mathbf{x}_T;\boldsymbol{\theta}_g)}[e^{-\mathcal{L}}] \tag{7}$$

Specifically, we use the REINFORCE algorithm [39] to train the on/off gate:

$$\nabla_{\boldsymbol{\theta}_g} J(\boldsymbol{\theta}_g) = E_{P(g|\mathbf{x}_T;\boldsymbol{\theta}_g)}[\nabla_{\boldsymbol{\theta}_g} logP(g|\mathbf{x}_T; \boldsymbol{\theta}_g)(e^{b-\mathcal{L}})] \tag{8}$$

where b is the moving average of previous losses, serving as the reinforcement baseline [39]. Based on [48], an empirical approximation of the above quantity is:

$$\nabla_{\boldsymbol{\theta}_g} J(\boldsymbol{\theta}_g) = \frac{1}{n} \sum_{k=1}^{n} [\nabla_{\boldsymbol{\theta}_g} logP(g|\mathbf{x}_T; \boldsymbol{\theta}_g)(e^{b-\mathcal{L}_k})] \tag{9}$$

where n is the number of different input datasets the gate samples (here, we take batches as sampled input datasets), and L_k is the cross entropy loss on the training dataset after the model is trained on kth sampled set. The training detail of the on/off gate is summarized in Algorithm 1.

4.3 Training Process

The above has explained how we formulate our input representation. Next, we introduce how to jointly train the gated VAE and the fake news detector together.

Suppose our training data is denoted as $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ and the learned latent variable is denoted as $\mathbf{z}$. $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ denote parameters of the encoder and the decoder, respectively. We want to estimate the true parameters of the generative model VAE. Therefore, the training objective is to maximize the (log) likelihood of training data, which can be expressed in Eq. 10 [19].

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \tag{10}$$

The first KL divergence term measures the closeness of the estimated distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ and the true distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})$) (which cannot be computed explicitly because we do not have knowledge of the true parameters). But since KL divergence is non-negative, the second term is called the *lower bound* on the marginal likelihood of $\mathbf{x}^{(i)}$, and can be rewritten as Eq. 11.

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbf{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \tag{11}$$

Combining Eq. 10 and Eq. 11, our VAE's training objective becomes minimizing the KL divergence of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ and $p_{\boldsymbol{\theta}}(\mathbf{z})$ and maximizing the expected log likelihood of $\mathbf{x}^{(i)}$ (equivalent to minimizing the reconstruction error).

Our fake news detector, D, consists of one fully connected layer followed by the softmax function. The input to D is the learned multimodal representation $\hat{\mathbf{z}}_{I+T}$ (see Sec. 4.1.2) and the output y_{pred} is $\{real, fake\}$. The loss function being used is cross entropy:

$$\mathcal{L}_{\mathcal{D}} = -\mathbf{E}_{y\sim Y}\left[y \log y_{pred} + (1-y)\log(1-y_{pred})\right] \tag{12}$$

To put everything together, we calculate the total loss by summing up the KL divergence, reconstruction errors and the cross entropy.

$$\mathcal{L} = \lambda_I rec_I + \lambda_T rec_T + \lambda_{kl} KLD + \lambda_D \mathcal{L}_D \tag{13}$$

Our experiments show that when setting all lambdas to 1, the model reaches the best performance. During training, we calculate gradient descent on $\mathcal{L}$ and back propagate it to update the model's parameters. The unimodal gate and multimodal gate's parameters are updated together with the VAE and fake news detector. The parameters of the RL gate is updated using Algorithm 1, separate from the training of the main model.

5 EXPERIMENTAL SETUP

We next present our experimental design to evaluate GatedVAE.

5.1 Datasets

We use two benchmark datasets to evaluate our approach.

5.1.1 Twitter Dataset [6]. This dataset is published by the *Verifying Multimedia Use* task at MediaEval. News posts are collected from Twitter. The labelled dataset consists of a development dataset (devset) and a test dataset (testset). The devset comprises of 6,225 real and 9,596 fake posts related to 17 events; the testset comprises of 1,107 and 1,121 posts related to 21 events. Posts in this dataset are associated with text and images/videos, and labelled with either *real* or *fake*. Following [35, 38], we preprocess the data by pairing each post's text with its first associated image, and removing posts that do not have an image, or the image file does not exist. After preprocessing, the devset contains 13,407 posts (real: 5,860, fake: 7,547), the test set contains 1,040 (real: 450, fake: 590) posts.

5.1.2 Weibo Dataset [16]. This dataset is collected from Weibo, the largest Chinese social media platform. Posts are crawled on the official rumor debunking system of Weibo. The original training and testing sets contain a number of tweets approximately with a ratio of 8:2. Posts in this dataset are associated with text and images, and labelled with either *real* or *fake*. We preprocess Weibo by pairing each post's text with its first associated image appearing in the image folder (otherwise the label distribution is extremely unbalanced), and removing posts whose image files do not exist. After preprocessing, the training set contains 6,154 (real: 2,807, fake: 3,347) posts, the test set contains 1,699 (real: 835, fake: 864) posts.

5.2 Baseline Methods

We compare our model's performance with the following SOTA baseline models:

BERT [10] is a pre-trained language model built upon transformer encoders. We adopt the BERT base uncased model pretrained in English and Chinese, followed by a fully connected layer (768 hidden units) as a classifier.

VGG-19 [33] is an image classification network built upon CNN layers. We adopt the VGG-19 pretrained on ImageNet dataset and change the output channel to 2 in the last fully connected layer.

EANN [38] is an end-to-end multimodal fake news detection framework that can derive event-invariant features and thus benefit the detection of fake news on newly arrived events.

MVAE [18] is a multimodal fake news classifier that uses a bimodal variational autoencoder coupled with a binary classifier to detect fake news.

SpotFake [35] is a multimodal framework that combines text features learned from language model and image features learned from vision model to detect fake news.

MCNN [43] is a multimodal framework for fake news detection that considers the consistency of multimodal data and captures the overall characteristics of social media information.

MCAN [41] is a multimodal framework for fake news detection. It is build upon layers of multimodal fusion with co-attention networks.

SAFE [47] is a similarity-aware fake news detection method which investigate multimodal information of news articles.

CAFE [9] is an ambiguity-aware multimodal fake news detection method that leverages the ambiguity between text and image to dynamically adjust the weight of unimodal and cross-modal features.

Note, for models (*i.e.* BERT, VGG-19, SpotFake, MCAN) that have publicly available code, we re-run their model with default configurations. For models (*i.e.* MCNN, SAFE, CAFE) that we fail to acquire complete source code for, and models (*i.e.* EANN, MVAE) whose reproduced performances are much lower than their reported results, for fair comparison, we use their original reported performance in comparison to our approach.

5.3 Implementation Details

We use CLIP [30] to obtain news image embedding $\mathbf{x}_I$ of size 512, and LASER [4] to obtain news text embedding $\mathbf{x}_T$ of size 1024. The latent variable $\mathbf{z}_{I+T}$'s size is set to 512. For both image and text encoder, we set the number of hidden units to 256. We set the dropout rate to 0.2 (experimented with {0.2, 0.5, 0.8}) and the batch size to 64. Similarly, in the image and text decoder, we set the number of hidden units to 512 and 1024 respectively, and the

dropout rate is set to 0.2. For the fake news classifier, the number of hidden units is set to 2 followed by the softmax function.

The model is trained using the Adam optimizer with weight decay set to 1e-5. In addition, the initial learning rate is 1e-4 and divided by 4 every 8 epochs. The model is trained for 30 epochs with early stopping.

For both the unmidoal and multimodal soft gates, we set the number of hidden units to 1. The RL-enabled on/off gate is implemented as a neural network consisting of two linear layers with hidden units set to 32 and 2 respectively. Sigmoid activation is applied after the first linear layer. The number of samples n at each training step is set to the batch size. The RL-enabled on/off gate is trained using Adam optimizer with weight decay set to 1e-8 and learning rate set to 1e-4.

6 RESULTS AND ANALYSIS

Based on the above experimental setup, we next evaluate GatedVAE, before investigating key factors related to the performance.

6.1 Experimental Results

Tab. 3 presents the evaluation results of GatedVAE, as well as the baseline approaches. We report the performance of our approach using the three proposed gates in Tab. 3.

On both datasets, our proposed GatedVAE outperforms the baseline models across most metrics. Notably, $\text{GatedVAE}_{\text{multi}}$ achieves the highest recall of 90.5% (10.6% improvement) on the Twitter dataset. In addition, $\text{GatedVAE}_{\text{uni}}$ attains the highest F1 and accuracy on the Twitter dataset: we improve upon the SOTA by 3.7% in F1 and 1.1% in accuracy. $\text{GatedVAE}_{\text{uni}}$ also outperforms the SOTA on the Weibo dataset. Here, we improve the F1, accuracy, precision, and recall by 2.4%, 2.5%, 1.8%, and 1.6%, respectively. Across all experiments, the highest precision is achieved by $\text{GatedVAE}_{\text{uni}}$ on the Weibo dataset (90.2%). Further, our model has superior performance in recall, suggesting that it is better in identifying a wider range of fake news. That is to say, if deployed in the wild, our model can better reduce the load on human content moderators. Last, the intra-model comparison shows that RL-enabled gate (Sec. 4.2.3) has the poorest performance on both datasets. This suggests that an RL-enabled gate will not be useful in practice.

To show that our approach is able to address the challenges identified in Sec. 3, we repeat the perturbation experiments. on $\text{GatedVAE}_{\text{uni}}$. The results are shown in Tab. 4. In contrast to previous SOTA (Sec. 3) which are insensitive to textual perturbations (Tab. 2), $\text{GatedVAE}_{\text{uni}}$ sees obvious performance (1.5%, 6.4%, 6.4% in Acc) change. We argue that this confirms our approach's efficacy in filtering out noise in textual inputs, while retaining the sensitivity to image perturbations.

The reason that our model's performance levels surpass the SOTA on most metrics are two-fold. First, the gate module in our multimodal VAE is effective in addressing noise in textual inputs, which helps the model to effectively fuse multimodal information. Second, batch normalization and dropout help reduce the model's overfitting on the training set. In the following ablation study, we design and conduct experiments to further verify the above reasoning.

Table 3: Comparison of our proposed approach with unimodal and multimodal baselines on test set. Numbers are reported in accuracy (Acc), Precision (Prec), Recall (Rec) and F1. The best scores are highlighted in bold.

Dataset	Model	Acc	Prec	Rec	F1
Twitter	BERT	0.563	0.565	0.682	0.618
	VGG-19	0.567	0.567	0.613	0.589
	EANN	0.715	0.822	0.638	0.719
	MVAE	0.745	0.745	0.748	0.744
	SpotFake	0.760	0.808	0.759	0.783
	MCNN	0.784	0.814	0.850	0.831
	MCAN	0.803	**0.863**	0.751	0.803
	SAFE	0.766	0.777	0.795	0.786
	CAFE	0.806	0.807	0.799	0.803
	$\text{GatedVAE}_{\text{uni}}$ (ours)	**0.817**	0.798	0.888	**0.840**
	$\text{GatedVAE}_{\text{multi}}$ (ours)	0.798	0.767	**0.905**	0.831
	$\text{GatedVAE}_{\text{RL}}$ (ours)	0.734	0.658	0.793	0.685
Weibo	BERT	0.799	0.777	0.829	0.802
	VGG-19	0.635	0.643	0.634	0.639
	EANN	0.827	0.847	0.812	0.829
	MVAE	0.824	0.830	0.822	0.823
	SpotFake	0.854	0.884	0.821	0.851
	MCNN	0.832	0.858	0.801	0.828
	MCAN	0.848	0.868	0.840	0.854
	SAFE	0.816	0.818	0.815	0.817
	CAFE	0.840	0.855	0.830	0.842
	$\text{GatedVAE}_{\text{uni}}$ (ours)	**0.879**	**0.902**	**0.856**	**0.878**
	$\text{GatedVAE}_{\text{multi}}$ (ours)	0.877	0.900	0.853	0.876
	$\text{GatedVAE}_{\text{RL}}$ (ours)	0.756	0.747	0.692	0.767

Table 4: Results of the perturbation experiments on $\text{GatedVAE}_{\text{uni}}$.

Original		P1		P2		P3	
Acc	F1	ΔAcc	ΔF1	ΔAcc	ΔF1	ΔAcc	ΔF1
0.817	0.840	↓0.015	↓0.004	↓0.064	↓0.021	↓0.064	↓0.021
P4		P5		P6		P7	
Acc	F1	ΔAcc	ΔF1	ΔAcc	ΔF1	ΔAcc	ΔF1
↓0.147	↓0.162	↓0.264	↓0.135	↓0.191	↓0.132	↓0.159	↓0.176

6.2 Ablation Study

We next perform an ablation study on the different modules in GatedVAE, demonstrating that our gating mechanism and methods to reduce overfitting are both important for multimodal fake news detection. Specifically, we ablate (denoted as -) BN, dropout and gated module successively, and replace (denoted as /) the gated module with a simple text preprocessing mechanism that removes hashtags, URLs, and emoticons from the text. The evaluation results are summarized in Table 5. The base GatedVAE is $\text{GatedVAE}_{\text{uni}}$, which performs the best (see Tab. 3) across all GatedVAE variants.

Table 5: Ablation study on the batch normalization (BN) layer, dropout and gated module of GatedVAE. Numbers are accuracy (Acc), Precision (Prec), Recall (Rec), F1.

Dataset	Method	Acc	Prec	Rec	F1
Twitter	(-) BN	0.757	0.775	0.807	0.791
	(-) dropout	0.781	0.776	0.862	0.817
	(-) uni-gate	0.790	0.776	0.886	0.827
	(/) preprocessing	0.758	0.956	0.846	0.798
	$\text{GatedVAE}_{\text{uni}}$	0.817	0.798	0.888	0.840
Weibo	(-) BN	0.872	0.869	0.881	0.875
	(-) dropout	0.872	0.884	0.862	0.873
	(-) uni-gate	0.871	0.877	0.869	0.873
	(/) preprocessing	0.835	0.927	0.734	0.819
	$\text{GatedVAE}_{\text{uni}}$	0.879	0.902	0.856	0.878

On the Twitter dataset, we observe that removing BN from the VAE encoders and decoders significantly decreases accuracy (6%). It further triggers a 2.3% – 8.1% decrease in precision, recall and F1, respectively. Likewise, removing the dropout operation causes a 2.7% decrease in accuracy (0.2% – 2.2% decrease in precision, recall and F1). Nevertheless, we find that the model's performance on the *training* set remains unaffected (still reaching 99%), suggesting that overfitting is a key issue in fake news detection. On the Weibo dataset, ablating BN and dropout causes a smaller change in our model's performance (within 1%). Although the precision drops by around 3% in these two settings, the increase in recall compensates for it. This causes F1 to drop by only 0.3%.

We observe similar degradation when removing the gated module from $\text{GatedVAE}_{\text{uni}}$: a 2.7% and 0.8% performance drop in accuracy on Twitter and Weibo, respectively. For all the other metrics, the performance drop ranges from 0.2% – 2.5%, except for a 1.3% improvement on recall on the Weibo dataset. Furthermore, if replacing the proposed gated module with a simple text preprocessing mechanism, we see an obvious performance decrease on both datasets.

The above results confirm the efficacy of BN, dropout operations and gated module. However, we note that in most cases, the techniques are *less* effective on the Weibo dataset. We conjecture that this is because the text is less noisy in the latter (when compared to Twitter). To briefly explore this, for both datasets, we calculate a set of exemplar features that we consider to be noise (see Tab. 1). Tab. 6 summarizes the results. Confirming our conjecture, we find that posts in the Twitter dataset *do* exhibit noisier characteristics: they comprise of fewer words, more hashtags, more URLs and more emojis. This indicates that the Twitter dataset indeed contains more noise and that GatedVAE can effectively deal with this challenge.

6.3 Interpreting the Gates Outputs

Previously, we have shown that introducing the gated module positively contributes to the model's performance. In this subsection, we further investigate the output score generated from the gated modules to better understand the modality importance.

Table 6: Measurement of noise level in Twitter and Weibo datasets.

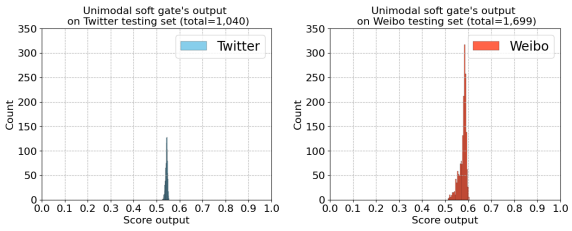
Feature (per post)	Twitter	Weibo
Avg text length	12.10	66.57
Avg # hashtags	1.394	0.435
Avg # URLs	1.084	0.002
Avg # emoticons	0.0581	0.0009

In Fig. 2, we plot the output score distribution of the gated modules on the two datasets.[2] Fig. 2(a) and 2(b) are generated from the unimodal gate. First, we note that both score ranges are very narrow, centering around 0.55; the score range on the Twitter testing set (in blue) is even narrower compared to the Weibo testing set (in red). Meanwhile, the former has a lower average score. We hypothesize that this discrepancy is resulting from different signal-to-noise ratios. Specifically, on Twitter, the textual inputs contain more noise (see Tab. 6); as such, the unimodal gate outputs lower scores to suppress the textual inputs. Fig. 2(c) and 2(d) present the scores generated from the multimodal soft gate. Recall that the score, g, represents the weight assigned to the textual inputs, while 1-g represents the weight assigned to the visual input (Sec. 4.2.3). In Fig. 2(c), the distribution is right skewed, indicating that the gate determines that the visual inputs weigh more than the textual inputs for the Twitter dataset. This is in line with our prior observations that text in Twitter contains more noise (see Tab. 6). In contrast, Fig. 2(d) reveals that the distribution of Weibo is close to a normal distribution with a mean around 0.5. This further confirms that the textual modality in the Weibo dataset is just as important as the visual modality. To explore this, we manually inspect a sample of 25 images from both Twitter and Weibo. We argue that this sample size offers a reasonable balance between representativity and the manual labour required. Our manual inspection reveals that, indeed, Twitter images tend to be narrower in event scope and more related to the news item at hand. In contrast, images within the Weibo dataset are *far* more diverse, with noticeable noise (*e.g.* unrelated images or cartoons). This again highlights the ability of GatedVAE to adaptively suppress different levels of noise contained in the news' text from social media dataset. We conjecture that the noise level difference in datasets is driven primarily by the two collecting methodologies employed (whereas the Twitter dataset covers a curated set of news events, the Weibo dataset covers a sample of any news event discussed during the measurement period).

6.4 Case Study

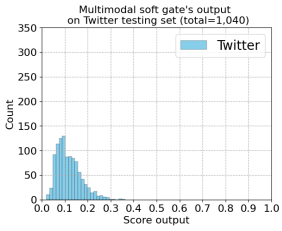
We investigate the scores output by gated modules to better highlight why news items receive different gated scores and discuss how scores assist in deciding the veracity of news items. Due to space constraints, we only discuss four representative news items in Fig. 3 and 4. We emphasize that these examples are only intended to offer context for our prior results.

[2]Note, we do not plot histograms for the on/off gate because we find that the outputs have very small variance, allowing the gate to block all textual inputs. This further shows that an on/off gate does not fit well for the two datasets we use, and totally removing textual inputs largely harms the model performance.

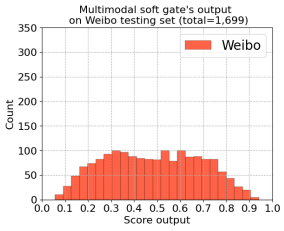


(a) Unimodal gate on Twitter

(b) Unimodal gate on Weibo



(c) Multimodal gate on Twitter



(d) Multimodal gate on Weibo

Figure 2: Histograms (bin=25) of scores output by the unimodal soft gate and multimodal soft gate. Blue stands for Twitter testing set, red for Weibo testing set.

(a) **News text:** Willkommen..!! http://t.co/yLvLNMr1pW **Uni:** 0.5245 **Multi:** 0.0673

(b) **News text:** Retweeted Cute Animal Pics (@CuteAnimalsBaby):The incredibly rare Black Lion, only a few of these exist... https://t.co/283cgcccKB **Uni:** 0.5463 **Multi:** 0.0885

Figure 3: Two Twitter fake news items respectively assigned with low and high (relatively) score by the unimodal gate, but assigned with similarly low score by the multimodal gate.

We first consider the unimodal gate. We find that GatedVAE$_{\text{uni}}$ tends to give higher scores to the post that has longer text and is more relevant to the news event it describes (Fig. 3(b) and 4(b)). In contrast, Fig. 3(a) and 4(a) receive relatively low scores from the unimodal gate; this is largely due to their short and less meaningful textual contents. This observation implies that the noise level on text does not vary a lot across different posts. However, it can serve as a factor guiding human moderators (when checking the text meaningfulness) to pinpoint the modality component that causes the news to be considered fake.

We also inspect the outputs of the multimodal gate, which considers the relative importance of both the textual and visual inputs. By comparing the selected four examples, we see that, if the image delivers a clear message, then it tends to receive a higher score; otherwise it gets a lower score. To highlight its operation, we use

(a) **News text:** Crying for a while **Uni:** 0.5255 **Multi:** 0.3518

(b) **News text:** ...we bought potato chips from 50 brands and give our ranking of chip flavours...(truncated) **Uni:** 0.6053 **Multi:** 0.1399

Figure 4: Two Weibo real news items respectively assigned with low and high (relatively) score by the unimodal/multimodal gate.

the news events in Fig. 3(a) and 3(b) as examples: the Paris Attack and the Black Lion. The semantic information contained within the news images provides enough information to determine whether the news is fake or not. This is likely why the textual inputs receives very low scores from the multimodal gate. In contrast, Fig. 4(a) and 4(b) conveys less comprehensible information (4(a) is even vaguer). In this scenario, the model credits higher weights to the textual inputs. Although the multimodal gate underperforms the unimodal gate, while deployed in the wild, it has value in directing human moderator's attention towards the corresponding modality that may cause the news to be fake.

We argue that this small set of examples helps shed light on our model's mechanism, which dynamically determines the weights assigned to different modalities based on its semantic importance. This also enables our model to obtain state-of-the-art performance.

7 CONCLUSION AND FUTURE WORK

This paper has proposed GatedVAE, a multimodal VAE fake news detector, coupled with a gated module, that can dynamically decide whether to pass through the noisy modality. Our experimental results show that GatedVAE not only obtains state-of-the-art performance (0.840 and 0.878 in terms of F1) on benchmark datasets, but also explains the decision making mechanism through the lens of the gated module. These advantages enable GatedVAE to better detect fake news and guide web content moderators to pinpoint the problematic modality input. We argue that such explanations are increasingly important in real-world scenarios.

We also note a set of limitations, which form the basis of our future work. First, we recognize that the RL-based gate is not as effective as the other proposed gates. We include it in order to make the solution space exploration more complete, and provide reference to future work (as it could be useful on other tasks). For example, it may be more suitable for temporal-based multimodal fake news detection where not every frame contributes equally (*e.g.* fake news video). Second, we note that our study has focused on two platforms, covering a relatively small set of events. In our future work, we plan to incorporate GatedVAE with contrastive learning and/or adversarial learning to enable domain adaptive multimodal fake news classification.

REFERENCES

[1] 2021. New Research Shows that 71% of Americans Now Get News Content via Social Platforms. https://www.socialmediatoday.com/news/new-research-shows-that-71-of-americans-now-get-news-content-via-social-pl/593255/.

[2] 2022. Social Media and News Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

[3] S. Afroz, M. Brennan, and R. Greenstadt. 2012. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *2012 IEEE Symposium on Security and Privacy*. 461–475.

[4] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610.

[5] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 549–556.

[6] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying Multimedia Use at MediaEval 2015. *MediaEval* 3, 3 (2015), 7.

[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[8] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 163–171.

[9] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.

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

[11] Yongping Du, Yang Liu, Zhi Peng, and Xingnan Jin. 2022. Gated attention fusion network for multimodal sentiment classification. *Knowledge-Based Systems* 240 (2022), 108107.

[12] Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2022. MMVAE at SemEval-2022 Task 5: A Multi-modal Multi-task VAE on Misogynous Meme Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, Seattle, United States, 700–710.

[13] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 943–951.

[14] Yi Han, Shanika Karunasekera, and Christopher Leckie. 2020. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316* (2020).

[15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[16] 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.

[17] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences* 9, 19 (2019), 4062.

[18] 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.

[19] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[20] Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. 2021. Multimodal Detection of Information Disorder from Social Media. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–4.

[21] Aysun Kocak, Erkut Erdem, and Aykut Erdem. 2021. A Gated Fusion Network for Dynamic Saliency Prediction. *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[22] Rina Kumari and Asif Ekbal. 2021. Amfb: attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications* 184 (2021), 115412.

[23] Fei-Fei Li, Justin Johnson, and Serena Yeung. 2019. Lecture notes of cs231n.

[24] Pengfei Liu, Kun Li, and Helen Meng. 2022. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition. *arXiv preprint arXiv:2201.06309* (2022).

[25] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).

[26] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909* (2020).

[27] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*. 3632–3640.

[28] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*. 3432–3441.

[29] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018).

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

[31] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. 2018. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3253–3261.

[32] Roberto Rey-de Castro and Herschel Rabitz. 2018. Targeted nonlinear adversarial perturbations in images and videos. *arXiv preprint arXiv:1809.00958* (2018).

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

[34] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In *Companion Proceedings of the Web Conference 2022*. 726–734.

[35] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.

[36] Thi Tran, Rohit Valecha, Paul Rad, and H Raghav Rao. 2019. An investigation of misinformation harms related to social media during humanitarian crises. In *International conference on secure knowledge management in artificial intelligence era*. Springer, 167–181.

[37] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. 2019. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2641–2650.

[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] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[40] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*. IEEE, 651–662.

[41] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.

[42] Zhiyuan Wu, Dechang Pi, Junfu Chen, Meng Xie, and Jianjun Cao. 2020. Rumor detection based on propagation graph neural network with attention mechanism. *Expert systems with applications* 158 (2020), 113595.

[43] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.

[44] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A Convolutional Approach for Misinformation Identification.. In *IJCAI*. 3901–3907.

[45] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. *arXiv preprint arXiv:2012.04233* (2020).

[46] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. 2018. Gated fusion network for joint image deblurring and super-resolution. *arXiv preprint arXiv:1807.10806* (2018).

[47] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: similarity-aware multi-modal fake news detection (2020). *Preprint. arXiv* 200304981 (2020), 2.

[48] Barret Zoph and Quoc V Le. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578* (2016).