MVAE: Multimodal Variational Autoencoder for Fake News Detection

Dhruv Khattar
International Institute of Information Technology
Hyderabad, India
dhruv.khattar@research.iiit.ac.in

Jaipal Singh Goud
International Institute of Information Technology
Hyderabad, India
jaipal.singh@research.iiit.ac.in

Manish Gupta*
International Institute of Information Technology
Hyderabad, India
manish.gupta@iiit.ac.in

Vasudeva Varma
International Institute of Information Technology
Hyderabad, India
vv@iiit.ac.in

ABSTRACT

In recent times, fake news and misinformation have had a disruptive and adverse impact on our lives. Given the prominence of microblogging networks as a source of news for most individuals, fake news now spreads at a faster pace and has a more profound impact than ever before. This makes detection of fake news an extremely important challenge. Fake news articles, just like genuine news articles, leverage multimedia content to manipulate user opinions but spread misinformation. A shortcoming of the current approaches for the detection of fake news is their inability to learn a shared representation of multimodal (textual + visual) information. We propose an end-to-end network, Multimodal Variational Autoencoder (MVAE), which uses a bimodal variational autoencoder coupled with a binary classifier for the task of fake news detection. The model consists of three main components, an encoder, a decoder and a fake news detector module. The variational autoencoder is capable of learning probabilistic latent variable models by optimizing a bound on the marginal likelihood of the observed data. The fake news detector then utilizes the multimodal representations obtained from the bimodal variational autoencoder to classify posts as fake or not. We conduct extensive experiments on two standard fake news datasets collected from popular microblogging websites: Weibo and Twitter. The experimental results show that across the two datasets, on average our model outperforms state-of-the-art methods by margins as large as ~6% in accuracy and ~5% in F_1 scores.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Neural networks**.

*Author is also a Principal Applied Researcher at Microsoft. (gmanish@microsoft.com)

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313552

KEYWORDS

Fake news detection, multimodal fusion, variational autoencoders, microblogs

ACM Reference Format:
Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313552

1 INTRODUCTION

The change to neoteric methods of news consumption in recent times has brought the issue of fake news and misinformation to the forefront of discussion. As thousands of new news articles proliferate social media networks every day, with each having neither a credibility nor a validation check, an ecosystem fuelled by misinformation (the inadvertent sharing of false information) and disinformation (the deliberate creation and sharing of information known to be false) has been established. Social media networks have enabled news articles to evolve from traditional text-only news to news with images and videos which can provide a better storytelling experience and has the ability to engage more readers. Recent fake news articles leverage this very change to visual context aided news. Fake news articles can now contain misrepresented, irrelevant and forged images to mislead readers. Peer-to-peer networks (primarily social media networks) allow fake news (propaganda) to be targeted at users who are more likely to accept and share a particular message. Fake news has been in the limelight in recent years for having an extensive negative effect on public events. A major turning point of realization was the 2016 U.S. presidential elections. It was believed that within the final three months leading up to the election, fake news favoring either of the two nominees was accepted and shared by more than 37 million times on Facebook [1]. This makes the task of detecting fake news a crucial one.

Figure 1 shows three examples of fake news from the Twitter dataset. Each tweet has certain textual content and an image associated with it. For the tweet on the left, both the image and text indicate that it is probably a fake news. In the tweet on the right, the image does not add substantial information but the text indicates that it may be a fake news. In the tweet in the middle, it is difficult to reach a conclusion from the text, but the morphed image suggests that it is possibly a fake news. This example reflects the hypothesis

Text: **Elephant carved out of solid rock. Magnificent!**

Text: **New species of fish found in Arkansas.**

Text: **Woman, 36, gives birth to 14 children from 14 different fathers**

Figure 1: Fake News examples from the Twitter dataset

that pairs of visual and textual information can give better insights into fake news detection.

On a conceptual level, the task of detecting fake news has undergone a variety of labels from misinformation to rumor. The target of our study is detection of news content that is fabricated and can be verified to be false. Rumor and fake news detection techniques range from traditional learning methods to deep learning models. Initial approaches [4] tried to detect fake news using only linguistic features extracted from the text content of news stories. Ma et al. [14] explored the possibility of representing tweets with deep neural networks by capturing temporal-linguistic features followed by Chen et al. [5] who built upon this by introducing attention mechanism into the RNNs.

Recent works [26] in the field of deep learning to detect fake news have shown performance improvements over traditional methods due to their enhanced ability to extract relevant features. Jin et al. [9] combine visual, textual and social context features, using an attention mechanism to make predictions about fake news. Wang et al. [26] use an additional event discriminator to learn common features shared among all the events with the goal of doing away with non-transferable event-specific features, and claim that they can handle novel and newly emerged events better. A shortcoming of the existing models is that they do not have any explicit objective function to discover *correlations across the modalities*.

Our work, inspired by the idea of autoencoders tries to learn shared representations in a multimodal setting. Kingma et al. [12] proposed that latent variable models act as a powerful approach to generatively model complicated distributions and brought forth the ideas of Variational Autoencoder (VAE). VAEs can learn probabilistic latent variable models by optimizing a bound on the marginal likelihood of the observed data. Overcoming the limitations of the current models, we propose a multimodal variational autoencoder capable of learning shared (visual + textual) representations, trained to discover correlations across modalities in tweets. The VAE is then coupled with a classifier to detect fake news.

To summarize, the contributions of our work are as follows:

- We propose a novel approach for classifying social media posts using only the content of the post, i.e., the text and the attached image.
- The proposed MVAE model uses a Multimodal Variational Autoencoder trained jointly with a Fake News Detector to detect if a post is fake or not.
- We extensively evaluate the performance of our model on two real-world datasets. The results reveal that our proposed model learns better multimodal features and outperforms the state-of-the-art multimodal fake news detection models.
- We show that our proposed model is able to discover correlations across the modalities and thus come up with better multimodal shared representations.

The rest of this paper is organized as follows: In Section 2, we discuss previous work on fake news detection and basics of autoencoders. In Section 3, we present our proposed model and its different components. In Section 4, we describe the datasets, baselines and provide the implementation details of our proposed model. Results and analysis are shown in Section 5, and we conclude with a brief summary in Section 6.

2 RELATED WORK

The task of fake news detection is similar to various other interesting challenges ranging from spam detection [29] to rumor detection [24] to satire detection [20]. As every individual may have their very own intuitive definition of such related ideas, each paper embraces its own definition. Following the previous work [21, 22], we specify that the target of our study is detecting news content that is fabricated and can be verified as false.

A few early studies tried to detect fake news based on linguistic features extracted from the text of news stories. Castillo et al. [4] employ a set of linguistic features such as special characters, sentiment positive/negative words, emojis, etc. to detect fake news. Popat et al. [19] use stance and language stylistic features like assertive verbs, discourse markers, etc. to assess the credibility of a claim. Context-free grammar rules were used to identify deception in Feng et al. [6]. Ma et al. [14] were the first to explore the possibility of representing tweets with deep neural networks by capturing temporal-linguistic features. Chen et al. [5] incorporated attention mechanism into recurrent neural networks (RNNs) to pool out distinct temporal-linguistic features with a particular focus.

The social connection established as a result of the interaction between users and tweets give birth to rich social context for posts. Wu et al. [28] infer embeddings of social media user profiles and leverage an LSTM network over propagation pathways to classify fake news. Liu et al. [13] model the propagation path of a news story as a multivariate time series and detect fake news through propagation path classification with a combination of RNNs and CNNs. Ma et al. [15] learn representations of a tweet using recursive neural models based on tree-structured neural networks. Jin et al. [8] used hand-crafted social context features such as the number of followers, retweets, etc. Most social context features are unstructured and usually involve an intensive amount of manual labour to collect. Also, they cannot provide sufficient information for newly emerged events.

Recent studies have shown that visual features (images) play a very important role in detecting fake news [27]. However, verifying the credibility of multimedia content on social media has received less amount of scrutiny. Extraction of basic features of attached images has been explored in [18]. However, these features are still hand-crafted and can hardly represent complex distributions of image content.

Deep neural networks have yielded immense success in learning image and textual representations. They have been successfully applied to various tasks including image captioning [10, 25], visual

question answering [2], and fake news detection [8, 26]. Jin et al. [8] proposed a model which extracts the visual, textual and social context features, and fuses them by attention mechanism. Wang et al. [26] learn event-invariant features using an adversarial network along with a multimodal feature extractor. However, both these models do not have any explicit objective to discover *correlations across the modalities*.

To overcome the limitations of the existing multimodal fake news detectors, we propose a multimodal variational autoencoder (MVAE) that learns a shared representation of both the modalities, text as well as image. The multimodal variational autoencoder is trained to reconstruct both the modalities from the learned shared representation and thus discovers *correlations across modalities*. We jointly train the multimodal variational autoencoder along with a classifier to detect fake news. Also, we use less information than our baselines to detect fake news i.e. we don't use any social or event related information.

3 THE PROPOSED MVAE MODEL

3.1 MVAE Overview

We propose a novel deep multimodal variational autoencoder (MVAE) to address the problem of fake news detection. The basic idea behind MVAE is to learn a unified representation of both the modalities of a tweet's content. The overall architecture of the proposed MVAE is illustrated in Figure 2. It has three main components:

- Encoder: It encodes the information from text and image into a latent vector.
- Decoder: It reconstructs back the original image and text from the latent vector.
- Fake News Detector: It uses the learned shared representation (latent vector) to predict if a news is fake or not.

3.2 Encoder

The inputs to the encoder are the text of the post as well as the image attached to it, and it outputs a shared representation of the features learnt from both the modalities. The encoder can be broken down into two sub-components: textual encoder and visual encoder.

3.2.1 Textual Encoder. The input to the textual encoder is the sequential list of words in the posts, $T = [T_1\ T_2\ \dots\ T_n]$, where n is the number of words in the text. Each word $T_i \in T$ in the text is represented as a word embedding vector. The embedding vector for each word is obtained with a deep network which is pre-trained on the given dataset in an unsupervised way.

To extract features from textual content, we use recurrent neural networks (RNNs) with Long-Short Term Memory (LSTM) cells. RNNs are a class of artificial neural networks which utilize sequential information and maintain history through their intermediate layers. A vanilla RNN has an internal state whose output at every time-step can be expressed in terms of the previous time step. However, it has been seen that vanilla RNNs suffer from a problem of vanishing and exploding gradients [7, 17]. This leads to the model learning inefficient dependencies between words that are a few steps apart. To overcome this problem, LSTM extends the basic RNN by storing information over long time periods by their use of memory units and efficient gating mechanisms.

Let $[h_1\ h_2\ \dots\ h_n]$ represent states of the LSTM and its state updates satisfy the following equations:

$$\left[f_t,\ i_t,\ o_t\right] = \sigma\left(Wh_{t-1} + Ux_t + b\right) \tag{1}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left(Wh_{t-1} + Ux_t + b\right) \tag{2}$$

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{3}$$

where, σ is the logistic sigmoid function, f_t, i_t, o_t represent the forget, input and output gates respectively , x_t denotes the input, h_t represents the hidden state at time t. The forget, input and output gates control the flow of information throughout the sequence. W, U are matrices which represent the weights and b are the biases associated with the connections.

We employ stacked bi-directional LSTM units to extract textual features. The final hidden state of LSTM is obtained by concatenating the forward and backward states. Finally, we pass the LSTM output through a fully connected layer (enc text fc) to get the textual features.

$$R_T = \phi\left(W_{tf}R_{lstm}\right) \tag{4}$$

where, R_{lstm} is the output of the LSTM, W_{tf} is the weight matrix of the fully connected layer and ϕ is the activation function used.

3.2.2 Visual Encoder. The input to the visual encoder is the image attached with the post, V. As seen in various visual understanding problems, image descriptors trained using convolutional neural networks (CNNs) over large amounts of data have proven to be very effective. The implicit learning of spatial layout and object semantics in the later layers of the network has contributed to the success of these features.

We employ the pre-trained network of VGG-19 architecture [23] trained over the ImageNet database and use the output of the fully-connected layer (FC7). During the joint training process, we freeze the parameters of the VGG network to avoid parameter explosion. Finally, we pass the VGG output through multiple fully connected layers (enc vis fc*) to get the same sized representation of the image as that of text.

$$R_V = \phi\left(W_{vf}R_{vgg}\right) \tag{5}$$

where, R_{vgg} is the feature representation obtained from the VGG-19, W_{vf} is the weight matrix of the fully connected layer and ϕ is the activation function used.

The textual feature representation R_T and the visual feature representation R_V are then concatenated and passed through a fully connected layer to form the shared representation. We then obtain two vectors μ and σ from the shared representation. They can be treated as the mean and variance respectively of the distribution of the shared representation. Furthermore, a random variable ϵ is sampled from a previous distribution (e.g., Gaussian distribution). The final reparameterized multimodal representation denoted by R_m can be calculated as follows:

$$R_m = \mu + \sigma \circ \epsilon \tag{6}$$

We denote the encoder as $G_{enc}(M,\ \theta_{enc})$, where θ_{enc} denotes all the parameters to be learned in encoder, and M represents the set of multimedia posts. Hence, the output of the encoder for a multimedia post, m, is the multimodal representation:

$$R_m = G_{enc}\left(m,\ \theta_{enc}\right) \tag{7}$$

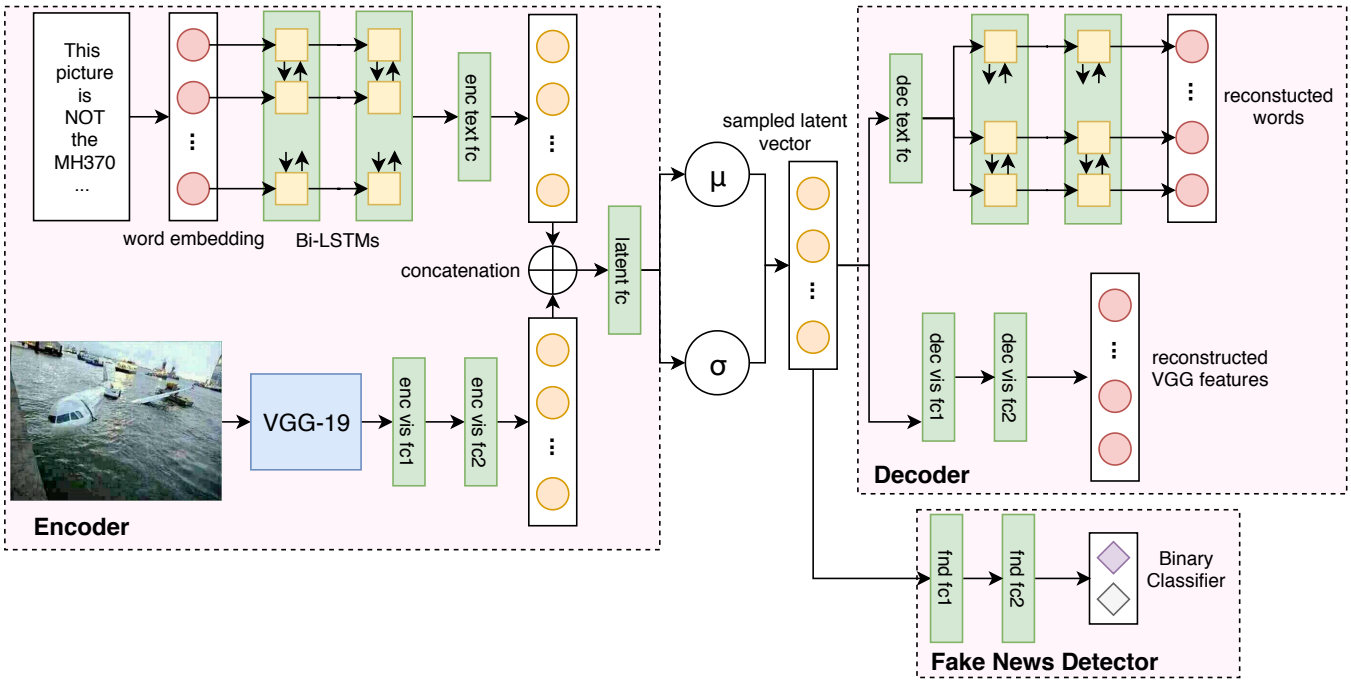

Figure 2: Network architecture of the proposed Multimodal Variational Autoencoder (MVAE). It has three components: Encoder, Decoder and Fake News Detector.

3.3 Decoder

The architecture of the decoder is similar to that of the encoder but inverted. The goal of the decoder is to reconstruct data from the sampled multimodal representation. Just like the encoder, it can be broken down into two sub-components: textual decoder and visual decoder.

3.3.1 Textual Decoder. The textual decoder takes as input the multimodal representation and reconstructs the words in the text. The multimodal representation is passed through a fully connected layer (dec text fc) to create inputs for the bi-directional LSTM. Then, we have similar stacked bi-directional LSTM units as used in the textual encoder sub-network. Finally, we pass the LSTM outputs through a time distributed fully connected layer with softmax activation to get the probability of each word in that time step.

3.3.2 Visual Decoder. The goal of the visual decoder is to reconstruct the VGG-19 features from the multimodal representation. The visual decoder sub-network is just the reverse of the visual encoder's corresponding sub-network. The multimodal representation is passed through multiple fully connected layers (dec vis fc*) to reconstruct the VGG-19 features.

We denote the decoder as $G_{dec}(R_m,\ \theta_{dec})$, where θ_{dec} denotes all the parameters in the decoder. Hence, the output of the decoder for a multimedia post, m, is a matrix of probability of every word for each position in the text, $\hat{t}_m$, and the reconstructed VGG-19 features of the image, $\hat{r}_{vgg_m}$.

$$\left(\hat{t}_m, \hat{r}_{vgg_m}\right) = G_{dec}\left(R_m,\ \theta_{dec}\right) \tag{8}$$

VAE models are trained by optimizing the sum of the reconstruction loss and the KL divergence loss. Therefore, we employ categorical cross-entropy loss for reconstruction of text and mean squared error for reconstruction of image features. The KL divergence between two probability distributions simply measures how much they diverge from each other. Minimizing the KL divergence here means optimizing the probability distribution parameters (μ and σ) to closely resemble that of the target distribution (Normal distribution). These can be calculated as follows:

$$\mathcal{L}_{rec_{vgg}} = \mathbb{E}_{m\sim M}\left[\frac{1}{n_v}\sum_{i=1}^{n_v}\left(\hat{r}_{vgg_m}^{(i)} - r_{vgg_m}^{(i)}\right)^2\right] \tag{9}$$

$$\mathcal{L}_{rec_t} = -\mathbb{E}_{m\sim M}\left[\sum_{i=1}^{n_t}\sum_{c=1}^{C}\mathbf{1}_{c=t_m^{(i)}}\log\hat{t}_m^{(i)}\right] \tag{10}$$

$$\mathcal{L}_{kl} = \frac{1}{2}\sum_{i=1}^{n_m}\left(\mu_i^2 + \sigma_i^2 - \log(\sigma_i) - 1\right) \tag{11}$$

where, M is the set of multimedia posts, n_v is the dimensionality of VGG-19 features, n_t is the number of words in text, n_m is the dimensionality of multimodal features, and C is the vocabulary size. We minimize the VAE loss by seeking optimal parameters $\hat{\theta}_{enc}$ and $\hat{\theta}_{dec}$ and this can be represented as follows.

$$\left(\theta_{enc}^*, \theta_{dec}^*\right) = \underset{\theta_{enc},\theta_{dec}}{\operatorname{argmin}}\left(\mathcal{L}_{rec_{vgg}} + \mathcal{L}_{rec_t} + \mathcal{L}_{kl}\right) \tag{12}$$

3.4 Fake News Detector

Fake News Detector takes multimodal representation as input and aims to classify the post as fake or not. It consists of multiple fully connected layers with corresponding activation functions. We

denote the fake news detector as $G_{fnd}(R_m,\ \theta_{fnd})$, where θ_{fnd} denotes all the parameters in the fake news detector. The output of the fake news detector for a multimedia post, m, is the probability of this post being a fake news.

$$\hat{y}_m = G_{fnd}\left(R_m,\ \theta_{fnd}\right) \tag{13}$$

We can view the value of $\hat{y}_m$ as a label 1 meaning the multimedia post m is fake, and 0 otherwise. In order to constrain the values between 0 and 1, we use the sigmoid logistic function. Hence, to calculate the classification loss, we employ cross-entropy as follows.

$$\mathcal{L}_{fnd}\left(\theta_{enc}, \theta_{fnd}\right) = -\mathbb{E}_{(m,y)\sim(M,Y)}\Big[y \log\left(\hat{y}_m\right) + \left(1-y\right)\log(1-\hat{y}_m)\Big] \tag{14}$$

where, M represents the set of multimedia posts and Y represents the set of ground truth labels. We minimize the classification loss by seeking the optimal parameters $\hat{\theta}_{fnd}$ and $\hat{\theta}_{enc}$, and this can be represented as follows.

$$\left(\theta^*_{enc}, \theta^*_{fnd}\right) = \underset{\theta_{enc}, \theta_{fnd}}{\operatorname{argmin}}\ \mathcal{L}_{fnd} \tag{15}$$

3.5 Putting it all together

The complete architecture of the proposed MVAE model is shown in Figure 2. The output of the encoder is fed to the decoder as well as the fake news detector. The decoder aims to reconstruct the data, while the fake news detector aims to classify the post as fake news or not. We jointly train the VAE and the fake news detector. Hence, the final loss can be written as follows.

$$\mathcal{L}_{final}\left(\theta_{enc}, \theta_{dec}, \theta_{fnd}\right) = \lambda_v \mathcal{L}_{rec_{vgg}} + \lambda_t \mathcal{L}_{rec_t} + \lambda_k \mathcal{L}_{kl} + \lambda_f \mathcal{L}_{fnd} \tag{16}$$

where, λs can be used to balance the individual terms of the loss function. In this paper, we simply set all the λs as 1. The optimal parameters can then be calculated by minimizing the final loss as follows.

$$\left(\theta^*_{enc}, \theta^*_{dec}, \theta^*_{fnd}\right) = \underset{\theta_{enc}, \theta_{dec}, \theta_{fnd}}{\operatorname{argmin}}\ \mathcal{L}_{final}\left(\theta_{enc}, \theta_{dec}, \theta_{fnd}\right) \tag{17}$$

4 EXPERIMENTS

In this section, we first describe the two social media datasets used in the experiments. We then discuss in brief the state-of-the-art fake news detection approaches along with some state-of-the-art language-vision models.

4.1 Datasets

Given the sparse availability of structured multimedia data, we make use of two standard datasets to evaluate our architecture for fake news detection. The two datasets consist of real social media information collected from Twitter and Weibo. To the best of our knowledge, these are the only available datasets that have paired image and textual information.

4.1.1 Twitter Dataset. As part of MediaEval [3], the Twitter dataset was released for Verifying Multimedia Use task. The task was aimed at detecting fake multimedia content on social media. The dataset consists of tweets (short messages posted on Twitter) and each tweet has textual content, image/video and social context information associated with it. The dataset has around 17000 unique tweets spanning over different events. The dataset is split into two parts: the development set (9000 fake news tweets, 6000 real news tweets) and the test set (2000 tweets). They are split in such a manner that the tweets have no overlapping events. Given our focus on the image and textual information, we filter out all tweets which have videos attached to them in the set. We use the development set for training and the test set for testing to keep the same data split scheme as the benchmark.

Table 1: Performance of MVAE vs other methods on two different datasets

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	**0.810**	0.498	0.617	0.584	0.759	0.660
	MVAE	**0.745**	0.801	**0.719**	**0.758**	**0.689**	**0.777**	**0.730**
Weibo	Textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	Visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.72	**0.889**	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	**0.824**	**0.854**	**0.769**	**0.809**	**0.802**	0.875	**0.837**

4.1.2 Weibo Dataset. The Weibo dataset, used in [8] for the task of fake news detection consists of data collected from Xinhua News Agency, an authoritative news source of China and Weibo, a Chinese microblogging website. The fake news collected from Weibo is obtained over a span of time ranging from May 2012 to January 2016 and is verified by Weibo's official rumour debunking system. The system encourages common users to report suspicious tweets on Weibo which are then examined by a committee of reputable users that classify the suspicious tweet as false or real. In accordance with previous work [14, 27], this system also acts as the authoritative source for collecting rumour news. The non-rumour tweets are tweets verified by the Xinhua News Agency. Preprocessing of the dataset is done using a method similar to that of [8]. Preliminary steps involve removal of duplicate images (using locality sensitive hashing) and low-quality images to ensure homogeneity across the dataset. The dataset is then split into training and testing sets with an approximate tweet ratio of 4:1 as in Jin et al. [8].

4.2 Experimental Settings

For word embeddings, we employ the distributed Word2Vec representation for words [16]. For the Twitter dataset, we have posts which are not in English, so we translated them into English to keep the data coherent. We follow standard text pre-processing for the Twitter dataset. For the Weibo dataset, text is in Chinese. Chinese text is written without spaces between words. We use the Stanford Word Segmenter for Chinese text tokenization into words.

We pre-train the Word2Vec model on the training dataset in an unsupervised fashion, with dimension size of 32.

For visual features, we use the output of the second to the last layer of a 19-layer VGGNet pre-trained on ImageNet set [23]. The dimensions of features obtained from VGG-19 is 4096. We do not fine tune the weights of the VGG, i.e., the weights of the VGG network are frozen.

The textual encoder consists of LSTMs with dimension size of hidden layers as 32. The fully connected layer used in the textual encoder is of size 32. The visual encoder consists of two fully connected layers of size 1024 and 32. The decoder consists of layers with the same dimensions as that of the encoder. The fake news detector has two fully connected layers of size 64 and 32.

We use a batch size of 128 instances in the training of the whole network. The model is trained for 300 epochs with a learning rate of 10^{-5} with an early stopping to report the results. We use the hyperbolic tangent function as the non-linear activation function. To prevent overfitting, we use L2-regularizer on the weights of our model. We experimented with a weight penalty of [0, 0.05, 0.1, 0.3, 0.5] and set it to 0.05 for the encoder and the decoder, and 0.3 for the fake news detector. To seek optimal parameters for our model, we use Adam [11] as the optimizer. We make the code publicly available[1].

4.3 Baselines

To validate the performance of the proposed multimodal variational autoencoder, we compare it with two categories of baseline models: single modality models and multimodal models.

4.3.1 Single Modality Models. MVAE leverages information from both visual and textual data to identify potentially fake news. As against such multimodal approaches, we also experimented with two unimodal models as described below.

- **Textual**: This model uses only textual information present in posts to classify them as fake or not. Each word is represented as a 32-dimensional vector. The word-embeddings are trained on the text content of the posts. Individual posts are then fed into a Bi-LSTM to extract textual features F_T. F_T is then fed into a 32-dimensional fully connected layer with a softmax function coupled with the Bi-LSTM that is responsible for making final predictions.
- **Visual**: The visual model uses only images from posts to classify them as fake or not. Images are fed into a pre-trained VGG-19 network with a fully-connected layer to extract visual features F_V. Similar to the textual model, the visual features F_V are then fed into a 32-dimensional fully connected layer for making the prediction.

4.3.2 Multimodal Models. Multimodal approaches utilize information from multiple modalities for the task of fake news classification.

- **VQA** [2]: Visual Question Answering aims to answer questions about given images. We adapted the Visual QA model which was originally designed for a multi-class classification task to our binary classification task. This is done by replacing the final multi-class layer with a binary-class layer. We use a one-layer LSTM with number of hidden units set to 32.
- **Neural Talk** [25]: The work of Vinyals et al. [25] in the domain of image captioning, proposes generation of natural language sentences describing an image using a deep recurrent framework. Following a structure similar to theirs, we obtain latent representations by averaging the output of the RNN at each time step as the joint representation of the image and text in tweets. These representations are then fed into a fully connected layer followed by an entropy loss layer to make predictions.
- **att-RNN** [8]: att-RNN uses attention mechanisms to combine textual, visual and social context features. In this end-to-end network, image features are incorporated into the joint representation of text and social context, obtained using an LSTM network. The neural attention from the outputs of the LSTM is an integral part for fusing the visual features. For a fair comparison, in our experiments, we remove the part dealing with social context information.
- **EANN** [26]: The Event Adversarial Neural Network (EANN) consists of three main components: the multimodal feature extractor, the fake news detector and the event discriminator. The multimodal feature extractor extracts textual and visual features from posts. It works with the fake news detector to learn the discriminative representation for detection of fake news. The event discriminator is responsible for removing any event-specific features. It is also possible to detect fake news using only two components, the multimodal feature extractor and the fake news detector. Hence, for a fair comparison, in our experiments, we work with a variant of EANN which does not include the event discriminator.

Note that fair comparisons with methods proposed in [13, 15] are not possible since they use addition information like tweet propagation data. Also, these methods are not developed for multimodal datasets. We report the accuracy, recall, precision, and F_1 scores of all the baselines along with our proposed model in Table 1.

5 RESULTS AND ANALYSIS

Table 1 shows the results of the baselines as well as our proposed method on both the datasets. We report the accuracy of our fake news detector as well as the precision, recall and F_1 score of our method for both fake news and real news. We can clearly see that our proposed method performs much better than the baselines.

On the Twitter dataset, among the single modality models, the visual model performs better than the textual model. This can be attributed to the fact that image features learnt with the help of VGG-19 have more shareable patterns to classify news as compared to textual features. Although visual features perform better than textual features, single modality models perform worse than the multimodal models.

Among multimodal models, att-RNN beats EANN which tells us that the attention mechanism can help in improving the performance of the model by considering the parts of the image which are related to the text. Our proposed model MVAE outperforms the baseline models by a huge margin and increases the accuracy from 66.4% to 74.5% and increases the F_1 scores from 66% to 73%.

On the Weibo dataset, we see similar trends in the results. The textual model beats the visual model in single modality models.

[1] https://github.com/dhruvkhattar/MVAE

Among multimodal methods, EANN and att-RNN which were proposed for this task perform better than Neural Talk and VQA. MVAE outperforms all the baselines and boosts the performance from 78.2% to 82.4% in terms of accuracy and shows an increase of ∼5% in F_1 scores, compared to the previous best baseline. This validates the effectiveness of our proposed method MVAE in detecting fake news on social media.

6 CONCLUSIONS

In this work, we explored the task of multimodal fake news detection. Overcoming the limitations of the current models, we tackle the challenge of learning correlations between modalities in tweets and to do so, propose a multimodal variational autoencoder that learns shared (visual + textual) representations to aid fake news detection. Our model consists of three main components, an encoder, a decoder and a fake news detector. Our proposed model, MVAE is trained by jointly learning the encoder, decoder and the fake news detector. The performance of our proposed architecture is evaluated on two real-world datasets. The MVAE model outperforms the current state-of-the-art architectures. In the future, we plan to extend MVAE using tweet propagation data and user characteristics.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. 31 (05 2017), 211–236.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E Middleton, Andreas Petlund, and Yiannis Kompatsiaris. [n. d.]. Verifying Multimedia Use at MediaEval 2016. ([n. d.]).

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

[5] Tong Chen, Lin Wu, Xue Li, Jun Zhang, Hongzhi Yin, and Yang Wang. 2017. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. *arXiv preprint arXiv:1704.05973* (2017).

[6] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[8] 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 2017 ACM on Multimedia Conference*. ACM, 795–816.

[9] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2017), 598–608.

[10] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

[13] Yang Liu and Yi fang Brook Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In *AAAI*.

[14] 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.. In *IJCAI*. 3818–3824.

[15] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1980–1989.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*. 1310–1318.

[18] Dong ping Tian et al. 2013. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering* 8, 4 (2013), 385–396.

[19] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2173–2178.

[20] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. 7–17.

[21] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[22] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

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

[24] Tetsuro Takahashi and Nobuyuki Igata. 2012. Rumor detection on twitter. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*. IEEE, 452–457.

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[26] 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*. ACM, 849–857.

[27] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.

[28] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 637–645.

[29] Yin Zhu, Xiao Wang, Erheng Zhong, Nathan Nan Liu, He Li, and Qiang Yang. 2012. Discovering Spammers in Social Networks.. In *AAAI*.