Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection

Shivangi Singhal
IIIT Delhi, India
shivangis@iiitd.ac.in

Tanisha Pandey
IIIT Delhi, India
tanisha17116@iiitd.ac.in

Saksham Mrig
IIIT Delhi, India
saksham19385@iiitd.ac.in

Rajiv Ratn Shah
IIIT Delhi, India
rajivratn@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT Hyderabad, India
pk.guru@iiit.ac.in

ABSTRACT

Recent years have witnessed a massive growth in the proliferation of fake news online. User-generated content is a blend of text and visual information leading to producing different variants of fake news. As a result, researchers started targeting multimodal methods for fake news detection. Existing methods capture high-level information from different modalities and jointly model them to decide. Given multiple input modalities, we hypothesize that not all modalities may be equally responsible for decision-making. Hence, this paper presents a novel architecture that effectively identifies and suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis. We also establish intra-modality relationship by extracting fine-grained image and text features. We conduct extensive experiments on real-world datasets to show that our approach outperforms the state-of-the-art by an average of **3.05%** and **4.525%** on accuracy and F1-score, respectively. We also release the code, implementation details, and model checkpoints for the community's interest.[1]

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**.

KEYWORDS

Multimodal Fake News Detection, Multiplicative Fusion, Fragment Embedding

ACM Reference Format:
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 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487553.3524650

[1]https://github.com/shiivangii/Leveraging-Intra-and-Inter-Modality-Relationship-for-Multimodal-Fake-News-Detection

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524650

Figure 1: An example of the tweet from the Twitter Dataset [4]. The corresponding text reads, 'Husband Gave His Unfaithful Ex-Wife Half Of Everything He Owned – Literally'. Our proposed intra-modality feature extractor curates the fine-grained salient representations for image and text, represented in the *blue* and *red* color, respectively.

1 INTRODUCTION

A husband divided his assets in half while settling the divorce case with her ex-wife. At a first read, the text might look believable. Now, when we read the same piece of information but with an image, shown in Figure 1, we might question the credibility of news. The image shows objects (i.e. car and laptop) cut into halves. We evaluated story's authenticity by analyzing information present in the different modalities associated with a news. The image proved to be a stronger signal than text in the example. A modality is strong when it can assign a high probability to the correct class. A higher probability implies a more informative signal and stronger confidence. The foundation of weak and strong modality is introduced in [20] and has been applied to various research domains [1, 22] to date. However, existing methods for multimodal fake news detection do not work on the principles of weak and strong modality [7, 16, 33, 34, 37, 43]. Instead, methods capture high-level information from different modalities and jointly model them to determine the authenticity of news. The feature extraction also occurs globally, ignoring the salient pixels containing meaningful information. For instance, Figure 1 highlights essential segments of the image and text containing details. However, current method of extracting visual features includes background information that might be unwanted. Similarly, there is a need to extract contextual dependencies for the textual features.

In this paper, we hypothesize that not all modalities play an equal role in the decision-making process on any particular sample. Therefore, we aim to design an architecture that utilizes a multiplicative multimodal method [20] to capture inter-modality relationship. The method suppresses the cost of a weaker modality by introducing a down-weight factor in the cross-entropy loss function. The down-weight factor associated with each modality highlights the average prediction power of the remaining modalities. So, if the other modality has higher confidence in predicting the correct class, cost associated with the current modality is suppressed and vice versa.

We also capture the intra-modality relationship. The idea is to generate fragments of a modality and then learn fine-grained salient representations from the fragments. For image modality, we perform bottom-up attention to extract the image patches [2]. The complex relationship between the patches is then encoded via self-attention mechanism [35]. The final visual representation is obtained by performing an average pooling operation over the fragment representations, resembling bag-of-visual-words model. We use a wordpiece tokenizer to generate text fragments for text modality. Taking inspiration from [8], we use a Transformer module, BERT, to extract contextual representations. The obtained embeddings are further passed through 1d-convolution neural network to extract the phrase-level information. The resultant text representation is obtained by passing intermediate learned representations via a fully connected layer.

Our contribution can be summarized as follows:

- **Capturing inter-modality relationship:** We present a novel architecture that uses a multiplicative multimodal method to capture the inter-modality relationship between modalities. Using the multiplicative multimodal method, we aim to leverage information from a more reliable modality than a less reliable one on a per-sample basis.
- **Capturing intra-modality relationship:** Our proposed method captures intra-modality relationship by extracting the fine-grained salient representations for image and text. The resultant feature vectors capture rich contextual dependencies present within its components.

Specifically, we aim to answer the following evaluation questions:

- **RQ1** Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships? (Section 5.3)
- **RQ2** How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection? (Section 5.4)
- **RQ3** Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample? (Section 5.5)

Experimental results on the two publicly available datasets show that our proposed method outperforms state-of-the-art methods by an average of **3.05%** and **4.525%** on the accuracy and macro F1-score, respectively.

2 RELATED WORK

Existing research has made several attempts to combat fake news. Such methods have primarily focused on mining textual clues in the form of lexical features [21], syntactic features [9], capturing writing styles [27] or extracting rhetorical structure [6]. In addition, few works have also explored other auxiliary features like, user comments [7], leveraging information from multiple news [14], exploring the propagation networks [29] or incorporating relational knowledge [39].

We have also observed a substantial growth toward methods that identify manipulations in the image [5, 12, 28]. An overview of the different image features that can be utilized for fake image detection is presented in [5]. Qi *et al.* [28] also proposed a novel method, Multi-domain Visual Neural Network (MVNN), that uses a CNN-based network to extract image features in the frequency domain and a multi-branch CNN-RNN model to extract visual features in the pixel domain. The final representations are obtained by fusing the feature representations of frequency and pixel domains.

We have recently witnessed a considerable evolution towards content-based multimodal fake news detection that combines text and corresponding images. This section revisits the works that emphasized incorporating images with the text for fake news detection.

2.1 Multimodal Fake News Detection

Jin *et al.* [13] made the first attempt towards multimodal fake news detection. Their paper proposed a recurrent neural network with an attention mechanism for fake news detection. It comprises of three sub-modules: first, sub-network uses RNN to combine text and social context features. The social context features are hashtags, mentions, retweets, and emotion polarity; Second, sub-network uses VGG19 pre-trained on the Imagenet database to generate representations for images present in tweets; Third, sub-network is a neural-level attention module that uses the output of RNN to align visual features. Yang *et al.*[41] made another attempt by designing a text and image information based Convolutional Neural Network (TI-CNN). The method extracts latent text and image features, represents them in a unified feature space, and then use learned features to identify fake news.

Another study by Wang *et al.* [37] proposed an event adversarial neural network for fake news detection. Core idea of the paper is to design a method that learns event-invariant features and preserve the shared features among all the events for fake news detection for newly emerged unseen events. The textual and visual features are extracted via Text-CNN and VGG19, respectively. The final representations are combined to form a multimodal feature vector utilized for fake news detection. In addition, the method uses an event discriminator to measure the dissimilarities among different events; it is a neural network that consists of two fully connected layers with corresponding activation functions. Khattar *et al.* [16] also came up with multimodal variational autoencoder for fake news detection. Model comprises of three components: (i) encoder, responsible for generating the shared representation of features learnt from both the modalities, (ii) decoder, responsible for reconstructing data from the sampled multimodal representation and,

(iii) fake news detector, that takes multimodal representation as input and classify the post as fake or not.

Another study attempts to detect fake news by leveraging spatial and frequency domain features from the image and textual features from the text present in a news [40]. Method uses multiple co-attention layers to learn the relationship between text and images. Visual features are first fused, followed by textual features; obtained fused representation from the last co-attention layer is used for fake news detection.

Recently, transformer-based language models have shown significant performance over traditional machine learning-based methods for fake news detection [26]. Singhal *et al.* came up with SpotFake [34] and Spotfake-plus [33]. that leverages textual information from the BERT and XLNet [42], respectively. Image features in both methods are extracted via VGG19 pre-trained on the Imagenet database.

All the works mentioned above have focused on multimodal fake news detection ignoring the relationship between textual and visual cues present in news articles. Zhou *et al.* [43] proposed a similarity-aware fake news detection method to investigate relationship between the extracted features across modalities. Text features are extracted via Text-CNN, and image feature generation is a two-step process. First, images are passed through the image2sentence model to generate a caption for the image. Generated text is then passed through Text-CNN to get the desired representations. A modified version of cosine similarity is used to establish a cross-modal relationship between the modalities. Another study by Singhal *et al.* [32] proposes a novel method that establishes a relationship between text and multiple images present in the news. The sequential information from the multiple visual cues is obtained by passing intermediate features obtained via VGG19 to the Bi-LSTM cells. Method uses BERT module for text feature extraction. A modified version of contrastive loss is used to establish the relationship between different news components.

Upon examining the related work, we find the following drawbacks, (i) Each method discussed before extracts visual information via Text-CNN or VGG19. Complete image is passed through the network to generate the representations. Image contains unwanted (redundant) information in the form of background that can be excluded, (ii) existing method combines different modalities to form multimodal feature vector. Such methods assume that both text and image modality play an equal role in determining the veracity of news. However, reports[2] show the existence of different versions of fake news due to manipulations performed in the different modalities. Hence there is a need to design a method that captures inter-modality relationship based on the modality contributing towards fake news.

3 PROBLEM FORMULATION

Assume we have a set of n news articles, $S=\{S_i^T, S_i^I\}_{i=1}^n$. Each news sample S_i consists of two elements, content (S_i^T) and the corresponding image (S_i^I). Our paper aims to capture the intra and inter-modality relationship to detect fake news. It is a binary classification task where S_i can be categorized as either fake *(y=0)* or real *(y=1)*.

[2]https://www.pagecentertraining.psu.edu/public-relations-ethics/introduction-to-the-ethical-implications-of-fake-news-for-pr-professionals/lesson-2-fake-news-content/types-of-fake-news/

Every content piece (S_i^T) comprises of k sentences, $\{S_i^{T_a}\}_{a=1}^k$. Each sentence $S_i^{T_a}$ is further tokenized into $\{w_{i1}, wi_{i2}, ..., w_{ik}\}$ subwords using a subword algorithm. The intra-modality relationship for text fragments is established by passing intermediate representations through multi-head self-attention layers. Finally, continuous representations, $\{z_s^{i:k}\}$ for each of the text piece are passed through a one dimensional convolution followed by a fully-connected layer to extract text representations. Similarly, every image (S_i^I) is segregated into a finite set of $\{m_i^1, m_i^2, ..., m_i^{36}\}$ fragments via a bottom-up attention module. The final image embeddings are obtained by performing average pooling over the continuous intermediate representations.

To correctly classify a news sample, we apply an efficient method that suppresses learning from the modality that independently incorrectly classified the sample. After, every modality $\{S_i^T, S_i^I\}$ makes its own independent decision with its modality-specific model, a multiplicative fusion method is utilized to mitigate the information gained from the weak modality by introducing a down-weight factor.

Problem Given a news sample, $S=\{S_i^T, S_i^I, Y\}$ where Y is the ground-truth label. Our goal is to design a novel architecture that (i) captures the intra-modality relationship via granular fragment representation and (ii) extracts the inter-modality relationship by inducing knowledge in the classification sub-module that tells which modality contributed towards fakeness. Such knowledge will also help readers understand the modality that contributed to the forgery.

4 METHODOLOGY

As shown in Figure 2, our proposed framework comprises of two components, an intra-modality relationship extractor and an inter-modality relationship extractor. Former gathers segment information from all the modalities independently; it derives global relationship among each fragment extracted for each modality. At the same time, latter is responsible for identifying strong and weak modalities via utilization of the multiplicative fusion. Next, we explain each component in detail.

4.1 Self Attention

Attention means to actively process a specific region in the environment while ignoring others. This section highlights how self-attention mechanism is used to model the intra-modality relationship for image and text fragments. We first provide an overview of the paradigm of attention function.

An attention module is a mapping function that takes in n inputs and returns n outputs. Every input comprises three representations: key, query, and value that interact and decide to whom they should pay more attention. The output is the aggregate of these interactions and attention scores.

Since we process the obtained set of image and text fragments independently in our approach, we can use self-attention, a particular case of the attention mechanism, to encode interaction between fragments of images or texts. In self-attention, all the three input representations, *i.e.* queries, keys and values are equal.

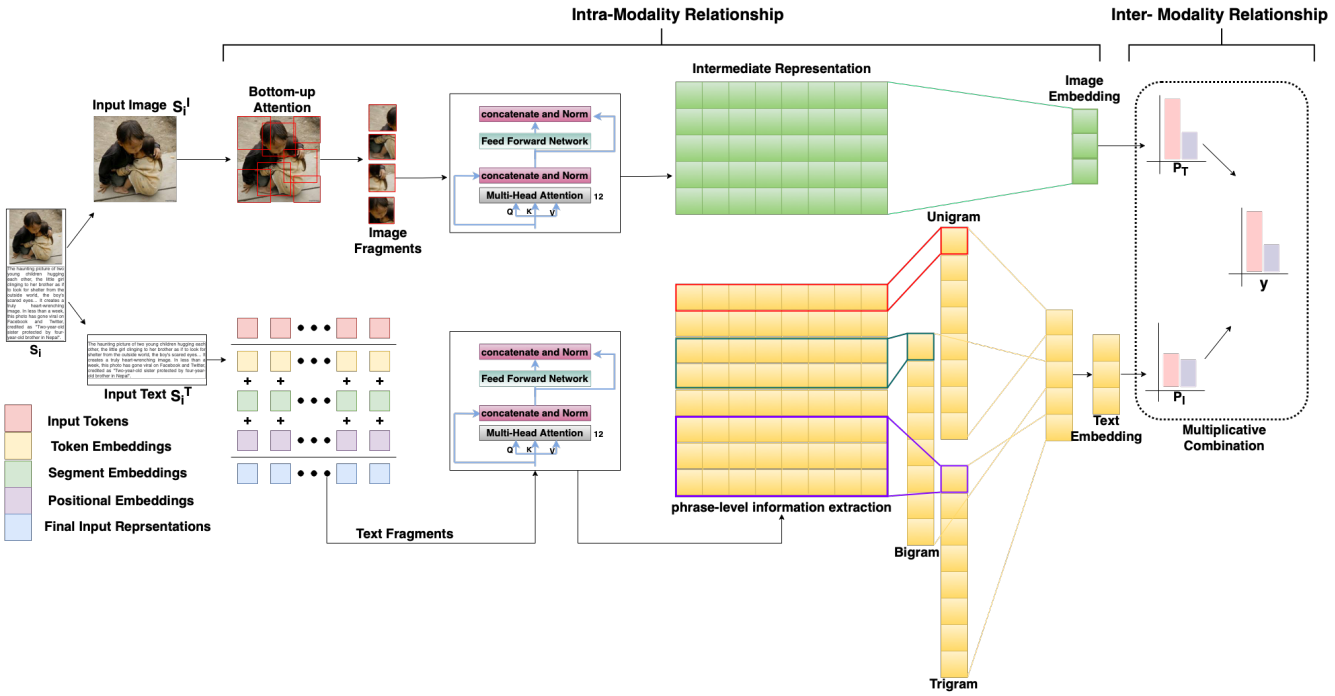

Figure 2: The high level diagram of our proposed model. It comprises two sub-modules. The intra-modality relationship extractor is responsible for extracting fragments and establishing relationships between them. The inter-modality relationship extractor is responsible for identifying the modality contributing to fakeness.

Taking inspiration from [35], we perform attention function via Transformers. A Transformer module embodies two sub-layers, multi-head self-attention sub-layer and position-wise feed-forward sub-layer. In the multi-head attention sub-module, attention mechanism runs through multiple times in parallel. Each attention head attends to a part of the sequence uniquely, and finally, all independent outcomes are combined and linearly reshaped to obtain the desired projection size.

For instance, let us assume that we have a finite set of fragments $f_1, f_2, \ldots f_b, f_b \in \mathbb{R}^{1xd}$, where b depicts the total number of fragments available for a modality and d is the representation size. Combining all these together resulted in a matrix $F = [f_1; \ldots f_b] \in \mathbb{R}^{bxd}$. Mathematically,

$$MultiHead = [head_1 \otimes, \ldots, head_i]W^O \quad (1)$$

$$head_i = attention(FW_i^Q, FW_i^K, FW_i^V)W^O \quad (2)$$

$$attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V \quad (3)$$

where Q, K, V are Query, Key-Value pair an W are all learnable parameter matrices. Next, the position-wise feed-forward submodule is applied on each fragment independently and identically to rearrange fragment embeddings in the preferred dimension:

$$FFN(g) = max(0, gW_1 + b_1)W_2 + b_2 \quad (4)$$

where $(g, b_1, b_2) \in \mathbb{R}^{1xd_g}$, $(W_1, W_2) \in \mathbb{R}^{d_g xd_g}$. At last, to promulgate position information to higher layers, residual connections followed by layer normalization is applied around each of the two sub-layers.

4.2 Text Embeddings

Each content piece (S_i^T) of a news article is segregated into sentences. A sentence s is represented as a sequence of WordPiece tokens $\{w_s^1, w_s^2, \ldots, w_s^k\}$, where w_s^k is aggregation of the token, position and segment representation for the *k-th* token present in a sentence s. Motivated by [8, 35], generated input sequence is passed to Transformers which is an attention based encoder-decoder type architecture. In our work, we make use of Transformer encoder that maps an input sequence of tokens $\{w_s^1, \ldots, w_s^k\}$ into an abstract continuous representation $\{z_s^1, \ldots, z_s^k\}$ that confines all the learned information of that input.

For instantiation, we adopt BERT (Bidirectional Encoder Representations from Transformers) architecture [8]. It is deeply bidirectional and looks at the words by jointly conditioning the left and right context in all layers. The context is pre-trained on Books Corpus and English Wikipedia to provide a richer understanding of language. BERT module is pre-trained with two unsupervised prediction tasks, next sentence prediction and masked language modelling. The former aims at predicting whether sentence A is the next sentence of B. In contrast, latter aims at masking some percentage of the input tokens at random and then predicting only those masked tokens.

Continuous representations obtained for each of the textual fragments, $Z_s = [z_s^1, \ldots, z_s^k]$ are then passed through a one dimensional convolution neural network to capture the hidden local context information of sequential features. Convolutional layer is used to

produce a feature map, $F_s = \{f_s^i\}_{i=1}^{k-h+1}$ from a sequence of continuous inputs $\{z_s^{i:(i+h-1)}\}_{i=1}^{k-h+1}$ via a filter w_s. Each local input is a group of h continuous words represented as,

$$f_s^i = \sigma(w_s \cdot z_s^{i:(i+h-1)} + b_s),$$
$$z_{i:(i+h-1)} = concat(z_i, z_{i+1}, \ldots, z_{i+h-1})$$

where $w_s, z_s^{i:(i+h-1)} \in \mathbb{R}^{hd}$, $b_s \in \mathbb{R}$ is a bias, σ is ReLU activation function and, w_s, b_s are the parameters learned within convolution neural network. After obtaining the convolution outputs, we apply max-pooling operation on the obtained feature map for dimensionality reduction, $\hat{f}_s = max\{f_s^i\}_{i=1}^{k-h+1}$. The text representations are then derived via $s = W_s\hat{f}_s + b_s$, where $\hat{f}_s \in \mathbb{R}^n; W_s \in \mathbb{R}^{qn}, b_s \in \mathbb{R}^q$. Specifically, $n \in \{1, 2, 3\}$ depicts three window sizes chosen for encapsulating the phrase level information at uni-gram, bi-gram and tri-gram level. Finally, the resultant text feature vector is obtained by passing s through a fully connected layer followed by L2 normalization.

4.3 Image Embeddings

Following [15, 19], for extracting the news image features, we primarily focused on extracting objects and other salient regions using a pre-trained detector. There are two reasons to abandon the classical method for image feature extraction. The embedding representations obtained from last pooling layer of VGG/CNN successfully preserve the spatial information. However, it might fail to capture the semantic relationship [11, 23, 36]. Classical approaches divide an image equally in spatial level, leading to redundant background information fragments. Filtering out unnecessary fragments demands additional computation and amendments in the algorithmic design.

Given an image I, we employ bottom-up attention model pretrained on Visual Genome [18] to extract a fixed-sized set of l image patches, $V = \{v_1, v_2, \ldots, v_l\}, v_l \in \mathbb{R}^d$ such that each image feature encodes a salient region and is represented by a pooled convolutional feature vector. The bottom-up attention module makes use of Faster R-CNN [30], a two-step object detection framework that identifies image patches belonging to certain classes and localizes them with bounding boxes. The first stage, identified as a Region Proposal Network, aims at predicting object bounds and objectness scores at each spatial position. In the second stage, the region of interest pooling is used to capture the feature map for each bounding box and classify image within the proposed region. Next, we add a position-wise fully connected layer to transform image features into a required dimension space for further processing *i.e.* $\{y_1, y_2, \ldots, y_l\}, y_l \in \mathbb{R}^{1xd}$. Finally, the resultant feature matrix is, $Y = [y_1; \ldots; y_l] \in \mathbb{R}^{lxd}$.

The intermediate representations obtained for each image fragment is then passed through a self-attention layer to capture complex relations among the image patches. With such a mechanism, each output fragment can attend to all input fragments, and the distance between each fragment is just one. The output obtained after passing through the multi-head self-attention module followed by layer normalization *(LN)* is, $O = [o_1; \ldots; o_l] \in \mathbb{R}^{lxd}$ where, $O = LN(Y + (MultiHead(Y))$. Then, the position-wise feed-forward and layer normalization is applied to get a set of continuous representations, $Z = \{z_i\}_{i=1}^{l}$, where $z_i = LN(o_i + FeedForward(o_i))$. Finally, the obtained image embeddings are condensed into a dense representations by performing average pooling followed by L2 normalization to procure the resultant image feature vector.

4.4 Multiplicative Multimodal Method

Our work aims to capture interaction among different modalities to better perform the task at hand. An intrinsic method to combine complementary information is to aggregate signals from different modalities and design learning models over concatenated features. Idea has been incorporated in numerous existing multimodal techniques including early and late fusion [10], hybrid fusion [3] and fusion methods enumerated from deep learning methods [24, 25, 38]. Intermediate representations are collated together and jointly modelled to decide in such methods. Such techniques are termed additive approaches due to the type of aggregation operation performed.

However, there are some practical constraints in integrating synergies across modalities using existing additive approaches. Additive methods assume that every modality is potentially helpful and is jointly combined to decide. Neural network models built on top of aggregated features cannot determine the quality of each modality and its contribution toward fake detection tasks on a per-sample basis. For instance, fake manipulations can be introduced by fabricating either text, images, or modalities. Given multiple input modalities, an ideal algorithm should be robust to noise from weak modalities and harvest relevant details from stronger modalities on a per sample basis. In this work, we perform the multiplicative multimodal method [20] that addresses the challenges mentioned above. Specifically, technique explicitly models that not all modalities contribute equally to any particular sample.

Let every modality present in a news sample make its own independent decision *i.e.* $P_T = [p_T^1; p_I^0]$ $P_I = [p_T^1; p_I^0]$, where P_T, P_I denotes the text and image predictions, respectively. Typical, additive combination would have resulted in,

$$l^y_{cross\_entropy} = -\sum_{i=1}^{M} \log(p_i^y)$$

where l^y is a class loss as it is part of the loss function associated with a particular class. To mitigate the challenges, we utilized a down-weight scaling factor,

$$q_i = [\prod_{j \neq i}(1 - p_j)]^{\beta/(m-1)}$$

where β is a hyper-parameter used to control the strength of down-weighting. The down-weight factor is responsible for suppressing the modality's predictive power that incorrectly classifies the sample. For instance, if p_i shows confident predictions for the correct class, down-weight factor will be a small value, suppressing cost for the other modalities ($j \neq i$). Intuitively, when current modality gives a favourable prediction, other modalities need not be equally helpful. Larger the value of down-weight factor, stronger the suppressing effect on that modality and vice versa.

Thus, when performing fake news classification task, we leverage benefits of extracting complementary information from the given piece of information using multiplicative method that have resulted

in the modification of loss function as,

$$l^{y}_{multiplicative} = -\sum_{i=1}^{M} q_i \cdot \log(p_i^y)$$

5 EXPERIMENTS

We first provide an overview of the dataset and baseline models, followed by a detailed investigation of the questions.

5.1 Datasets

We use two publicly available datasets to perform multimodal fake news detection,

- MediaEval Benchmark Dataset: The dataset is released as a part of the Verifying Multimedia Use task that took place as a part of MediaEval Benchmark in 2015 [4]. It comprises of 16,521 unique tweets with corresponding images. The dataset was curated around widely known 11 real-world events. There are 12,740 tweets in the training partition divided into 7,032 fake tweets and 5,008 real tweets. The testing partition comprises 2,564 fake tweets and 1,217 real tweets.
- Weibo Dataset: The dataset is introduced in [13]. The fake posts are collected from the official debunking system of Weibo from May 2012-January 2016. The tweets verified by Xinhua News Agency, an authoritative news agency in China, were considered for real posts. The dataset comprises 4,749 fake posts and 4,779 real posts partitioned into an 8:2 training and testing ratio.

5.2 Baselines

We compare our proposed methodology with a representative list of state-of-the-art multimodal fake news detection algorithms listed as follows:

- Text-CNN [17]: It is a deep learning algorithm that is capable of performing text classification. The algorithm uses a series of 1D convolutions and pooling layers to establish a semantic relationship between a text's words.
- BERT [8]: It is a transformer-based machine learning technique capable of extracting contextual meaning from the text. We used a version of BERT pre-trained on Wikipedia and Brown Corpus.
- VGG-19 [31]: It is a deep convolutional neural network that consists of 19 layers. It is used for classifying images. We used a version of the Vgg-19 network pre-trained on the ImageNet database.
- EANN [37]: It is an end-to-end framework that aims to capture event invariant features for fake news detection. The method extracts text and image features by employing Text-CNN [17] and pre-trained VGG19 [31] networks. The prime motivation to keep an event discriminator is to exclude event-specific features and keep shared features among events to better classify a fake sample on a newly emerged event.
- MVAE [16]: The algorithm seeks to establish correlation across modalities by designing a multimodal variational autoencoder. The module reconstructs representations of both modalities from the learned shared feature vector. This module is used in tandem with the classification module to detect fake news. The textual information is extracted via Bi-LSTMs and image features via VGG-19 pre-trained on the ImageNet dataset [31].
- SpotFake [34]: The algorithm leverages the power of language models to extract contextual text information [8]. The image feature is generated from the pre-trained VGG-19 network. The features obtained from both modalities are fused in an additive manner to build the desired news representation.
- Proposed w/o Text: It is a variant of the proposed method when using only visual information.
- Proposed w/o Image: It is a variant of the proposed method when using only textual information.
- Proposed w/o multiplicative fusion: It is a variant of the proposed method that fuses information from both modalities in an additive manner. Taking cues from the previous multimodal approaches [16, 34, 37], we used the late fusion strategy to perform the desired task.

5.3 Multimodal Fake News Detection RQ1

The question aims to examine the performance of the proposed model with the existing state-of-the-art models described in Section 5.2. The results are presented in Table 1 and 2. We used two unimodal text-based baselines comprising deep learning (Text-CNN) and transformer (BERT) based techniques. VGG19 is used as a unimodal image-based baseline. We also used several multimodal fake news detection SOTA methods (EANN, MVAE, SpotFake) for a fair comparison.

Results shown in the Table 1 and 2 indicate that our proposed method outperforms the baselines on accuracy and F1-score for Twitter and Weibo, respectively. SpotFake [34] is the strongest baseline on multimodal fake news detection, and our proposed method outperforms it by a fair margin of an average of **3.05%** and **4.525%** on the accuracy and F1-score, respectively.

Table 1: Comparison of our proposed model with unimodal text†, image$^{\mp}$ and multimodal‡ fake news detection baselines on the Twitter Dataset. Our proposed model beats the strongest baseline, SpotFake by 5.4% and 4.0% on accuracy and F1-score, respectively.

	MediaEval Benchmark Dataset			
Baselines	Acc	Prec.	Rec.	F1
Text-CNN†	0.614	0.599	0.612	0.594
BERT†	0.607	0.595	0.601	0.594
VGG-19$^{\mp}$	0.558	0.572	0.573	0.558
EANN‡	0.648	0.697	0.630	0.634
MVAE‡	0.745	0.745	0.748	0.744
SpotFake‡	0.777	0.791	0.753	0.760
Proposed	**0.831**	**0.836**	**0.832**	**0.830**

Figure 3: Different variants of fake news detected by our proposed model.

Table 2: Comparison of our proposed model with unimodal text†, image$^{\mp}$ and multimodal‡ fake news detection baselines on the Weibo dataset. Our proposed model beats the strongest baseline, SpotFake by 0.77% and 1.2% on accuracy and F1-score, respectively.

	Weibo Dataset			
Baselines	Acc	Prec.	Rec.	F1
Text-CNN†	0.794	0.791	0.800	0.792
BERT†	0.861	0.860	0.870	0.859
VGG-19$^{\mp}$	0.654	0.502	0.502	0.501
EANN‡	0.782	0.790	0.780	0.778
MVAE‡	0.824	0.830	0.822	0.823
SpotFake‡	0.8923	0.874	0.810	0.835
Proposed	**0.900**	**0.882**	**0.823**	**0.847**

Table 3: Comparison of our proposed model with its different variants.

	Variants	w/o Text	w/o Image	w/o Multiplicative	Proposed
Twitter	Acc	0.703	0.626	0.813	**0.831**
	Prec.	0.707	0.622	0.814	**0.836**
	Rec.	0.707	0.621	0.812	**0.832**
	F1	0.705	0.621	0.812	**0.830**
Weibo	Acc	0.736	0.794	0.873	**0.900**
	Prec.	0.608	0.802	0.824	**0.882**
	Rec.	0.588	0.791	0.815	**0.823**
	F1	0.595	0.791	0.820	**0.847**

5.4 Ablation Studies RQ2

The RQ2 aims to measure how effective the extracted fragments and self-attention representations are in improving multimodal fake news detection. To answer the question, we compare proposed model with its different variants. The question aims to establish effectiveness of each sub-module in the method. The results are shown in Table 3. For instance, *Proposed w/o Multiplicative* measures the effectiveness of the fusion strategy used in the paper. On average, we encounter a drop of **1.8%** and **2.7%** on the accuracy and F1-score, respectively, on removing the multiplicative fusion module. Similarly, to examine the effectiveness of the extracted fragments for the text and image modality, we evaluate the performance of the *Proposed w/o Text* and *Proposed w/o Image*, respectively.

5.5 Case Study RQ3

We perform a qualitative analysis of the proposed method to validate its efficacy in identifying the modality that easily recognises falsification in a particular news sample.

We took a random subset from the test set to examine the quality of the obtained q score. A few case studies are shown in Figure 3. Moreover, to validate the inferences, we also cross-examine the results with human interference in the loop. Following are the observations:

- Figure 3 (a) is a typical case of **False Context** where truthful information (Eiffel Tower lit up) is shared with the false contextual information (Barbaric attacks in Lahore). Sources[3] claim the information to be false, with no connection between Paris and Pakistan.
- Figure 3 (b) is a classic example of **Fabricated Content** where the content is created to deceive or do harm. Both text and image provide strong confidence in detecting the veracity. Hence, model's down-weight factor (q) assigned to the text and image is 0.7 and 0.4, respectively.
- Figure 3 (c) depicts an image of a girl claiming that she is selling chewing gum on the streets of Jordan. The story is an instance of **False Connection**. It is a case where no truth is established between the content's actual headline, image, or caption. On a closer inspection, we observe a happy emotion depicted in the image that is irrelevant to the war-like situation. Moreover, our model also shows stronger confidence in the Image modality by assigning a q score of 0.03 and 0.8695 to text and image, respectively.

[3]https://www.scoopwhoop.com/The-Photo-Showing-Eiffel-Tower-Lit-Up-In-Green-Is-Not-For-Victims-Of-Lahore-Blast/

- Figure 3 (d) highlights a clear case of a doctored photo presented with genuine information to deceive the readers. Since imaging modality shows more substantial confidence towards prediction, the model's performance highlights same. The example is a variant of fake news, often termed as **Manipulated Content**.

6 CONCLUSION

This paper presents a novel framework that leverages intra and inter modality relationships for multimodal fake news detection. Our proposed method comprises of two sub-modules. The first sub-module, intra-modality feature extractor, is responsible for extracting fine-grained salient image and text features. The final representations for text is obtained by passing raw text via BERT followed by Text-CNN. The image fragments are obtained via bottom-up attention, then passed through pooling layer to get the final representations. The second sub-module, inter-modality relationship extractor, fuses multimodal features multiplicatively. Such a fusion method can identify the modality that presented more substantial confidence towards fake news detection. Extensive experiments on two publicly available datasets demonstrate our proposed method's effectiveness.

ACKNOWLEDGMENTS

Shivangi Singhal is supported by TCS Research Scholar Program. Rajiv Ratn Shah is partly supported by the Infosys Center for AI at IIIT Delhi. We also thank Hitkul Jangra, Mehul Arora, Ritwik Mishra, Avinash Tulasi and Saurabh Gupta, members of the Precog Research Lab at IIIT-Hyderabad, for the valuable discussions.

REFERENCES

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14679–14689.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6077–6086.

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 345–379.

[4] 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.

[5] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 141–161.

[6] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology* (2015), 1–4.

[7] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 41–48.

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

[9] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic Stylometry for Deception Detection. In *Association for Computational Linguistics (ACL)*. 171–175.

[10] Hatice Gunes and Massimo Piccardi. 2005. Affect recognition from face and body: early fusion vs. late fusion. In *IEEE international conference on systems, man and cybernetics*. 3437–3443.

[11] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2310–2318.

[12] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 101–117.

[13] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. 795–816.

[14] Zhezhou Kang, Yanan Cao, Yanmin Shang, Tao Liang, Hengzhu Tang, and Lingling Tong. 2021. Fake News Detection with Heterogenous Deep Graph Convolutional Network.. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 408–420.

[15] 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.

[16] 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 (WWW)*. 2915–2921.

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision* (2017), 32–73.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[20] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint* (2018).

[21] David M Markowitz and Jeffrey T Hancock. 2014. Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one* 9, 8 (2014), e105937.

[22] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 1359–1367.

[23] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *IEEE conference on computer vision and pattern recognition (CVPR)*. 299–307.

[24] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. 2016. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016), 1692–1706.

[25] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML)*.

[26] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. *The Surprising Performance of Simple Baselines for Misinformation Detection.* Association for Computing Machinery, New York, NY, USA, 3432–3441. https://doi.org/10.1145/3442381.3450111

[27] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Association for Computational Linguistics (ACL)*. 231–240.

[28] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. 518–527. https://doi.org/10.1109/ICDM.2019.00062

[29] Bhavtosh Rath, Xavier Morales, and Jaideep Srivastava. 2021. SCARLET: Explainable Attention based Graph Neural Network for Fake News spreader prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems (NeurIPS)* (2015), 91–99.

[31] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, Yoshua Bengio and Yann LeCun (Eds.).

[32] Shivangi Singhal, Mudit Dhawan, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Inter-Modality Discordance for Multimodal Fake News Detection. In *ACM Multimedia Asia* (Gold Coast, Australia) *(MMAsia '21)*. Association for Computing Machinery, New York, NY, USA, Article 33, 7 pages. https://doi.org/10.1145/3469877.3490614

[33] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning. *Proceedings of the AAAI Conference* (2020), 13915–13916.

[34] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *IEEE International Conference on Multimedia Big Data (BigMM)*. 39–47.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[36] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. 2018. Joint Global and Co-Attentive Representation Learning for Image-Sentence Retrieval. In *ACM International Conference on Multimedia (ACMMM)*. 1398–1406.

[37] 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 *ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD)*. 849–857.

[38] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2362–2365.

[39] Kun Wu, Xu Yuan, and Yue Ning. 2021. Incorporating Relational Knowledge in Explainable Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

[40] 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.

[41] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749* (2018).

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[43] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.