CAMFeND: Credibility-Aware Multimodal Fake News Detection with Rotational Attention

Nidhi Gupta, Qinghua Li, and Lu Zhang
Electrical Engineering & Computer Science
University of Arkansas
Fayetteville, AR, USA
Email: {nidhig, qinghual, lz006}@uark.edu

***Abstract*—In the evolving digital landscape, fake news is a significant challenge, influencing public perception and decision-making. Traditional detection approaches focus on single-modal data or simple multimodal fusion, often overlooking deeper interactions and news credibility. We propose a novel model addressing these limitations by introducing rotational attention and news domain information as a feature. Unlike static attention mechanisms, our rotational attention dynamically shifts query, key, and value roles across text and image inputs, enabling richer cross-modal interaction. Incorporating news domain information further enhances the model's reliability by associating news posts with top domains extracted from Google search results, reducing false detections. This approach assesses both the content and the broader web context in which the news is discussed. Our model outperforms existing state-of-the-art methods by providing deeper, layered multimodal integration and domain information analysis, resulting in a more robust and adaptive fake news detection system.**

I. Introduction

In today's digital age, distinguishing between true and false information has become increasingly challenging. Many sources disseminate misleading or entirely fabricated content, undermining trust in reliable news outlets. For instance, high-profile incidents like the false reports of a deadly attack on a French satirical weekly, supposedly resulting in ten fatalities, and the fabricated story of a tragic shooting of a Canadian soldier in Ottawa (Figure 1), highlight the profound impact of fake news on public beliefs. These examples underscore the urgent need for advanced methods to analyze and verify the truthfulness of news. Developing state-of-the-art fake news detection technologies is essential for preserving the reliability of information sources and enhancing public understanding.

10 dead as shots fired at French satirical weekly

Canadian soldier shot in Ottawa a reservist from Hamilton

Fig. 1. Illustrations of fake news stories sourced from the Pheme [1] dataset.

Early detection approaches [2]–[5] primarily relied on machine learning techniques with manually crafted features from text and social context. Subsequent advancements introduced models designed to capture local dependencies in textual content by employing convolution-based methods [6]. Other approaches focused on modeling sequential information using recurrent structures [7], [8]. More recently, transformer-based methods have achieved significant progress by leveraging attention mechanisms to uncover deep semantic relationships within textual data [9]. These text-centric approaches fail to incorporate visual and multimodal clues, which are vital for detecting deceitful content. Recent research in multimodal fake news detection has emphasized the importance of integrating diverse sources of information. For example, [10] leverages latent representations for multimedia posts, while [11] combines BERT and VGG-19 features to enhance detection accuracy. Additionally, [12] addresses cross-modal inconsistencies, and [13] integrates features from various sources. However, two major limitations persist:

- **Limited Cross-Modal Interaction:** Many existing models struggle to capture the complex inter-modal relationships necessary for effective fake news detection. Approaches such as [14] with co-attention and [15] with relationship-aware attention rely on static feature alignment, assuming fixed interactions between modalities. This rigid approach fails to account for the evolving and dynamic relationships between text and image features, which are crucial for detecting fake news.
- **Neglect of News Domain Credibility:** Most models overlook the credibility of news domains as a feature, focusing solely on content analysis. This omission leaves the models vulnerable to misinformation from unverified or unreliable sources. Incorporating domain credibility is essential for filtering unreliable content and improving classification accuracy.

To address these limitations, we propose a novel fake news detection framework with two key components:

- **Rotational Attention Mechanism:** Traditional attention mechanisms, including co-attention and self-attention, rely on static roles for query (Q), key (K), and value (V) between text and image embeddings. While effective, this static role assignment may overlook intricate cross-modal dependencies, particularly in scenarios where the two modalities provide complementary or conflicting cues.

We propose a novel rotational attention mechanism which dynamically rotates the roles of Q, K, and V across layers, ensuring a more symmetric and comprehensive interaction, enabling each modality to influence and be influenced by the others from multiple directional perspectives. This richer, more nuanced interaction enhances the model's ability to resolve modality conflicts, such as when text and images convey contradictory information.

- **News Domain as a Credibility Feature:** We incorporate news domain information as a feature to address the issue of source credibility. Using Google's custom search API, we extract the top domains (e.g., bbc.com, time.com) based on the news text keywords. This contextual information provides insight into how a news topic is discussed across reliable and unreliable sources, enabling the model to filter out misinformation more effectively. By integrating domain credibility into the detection process, the model achieves greater robustness and accuracy.

Our framework surpasses previous multimodal fake news detection approaches by achieving better performance on benchmark datasets while maintaining lower complexity. By addressing the above limitations, our method offers a more robust and efficient solution to fake news detection. By dynamically rotating the roles of query, key, and value across modalities, the model processes multimodal data in multiple ways, ensuring balanced contributions from text, image, and domain-level information. Meanwhile, this mechanism enables simpler capture of diverse data representations, which enhances the model's effectiveness in detecting fake news.

We have conducted extensive empirical evaluations using Pheme [1] and Twitter [16] datasets. The results demonstrate significant improvements in performance across all baselines on the Twitter and Pheme datasets, validating the effectiveness of our proposed framework. Furthermore, an ablation study confirms that both the rotational attention mechanism and the incorporation of news domain credibility are critical to the model's superior performance, as their combined contributions drive the enhanced accuracy and robustness of our multimodal fake news detection solution, addressing a pressing societal issue.

II. Related Work

A. Single-Modal Approaches

Research on single-modal approaches to fake news detection initially focused on social and textual feature analysis. Early work such as [2], [17], [4] explored credibility through Twitter metadata, user behavior, writing style, and propagation patterns but were limited by surface-level analysis and lacked deeper content understanding. Similarly, approaches like [5], [18] leveraged time-series data and propagation structures but neglected content-based insights and semantic meaning.

With the rise of neural networks, RNN and CNN-based models such as [7], [19] improved feature-based and sequential analysis but struggled with long-term dependencies and contextual depth. They incorporated multi-domain elements and advanced text embeddings [20]–[22], while failed to capture dynamic feature interactions and struggled with ambiguity in generation-based models. Graph-based approaches (e.g., [23]) have also been proposed to improve rumor detection using graph convolutional networks; however, they still lack full multimodal feature integration.

B. Multimodal Approaches

Early multimodal fake news detection models integrated textual and visual data for better accuracy but lacked dynamic feature interactions. Event-invariant features, latent representations, and pre-trained models have been explored in prior works such as [24], [10], [11]. However, these approaches collectively struggled with event-specific variations, handling multimodal conflicts, and reliance on static features, which limited their adaptability. Cross-modal similarity has been a focus of prior research, such as the work in SAFE [25], but these approaches missed deeper semantic integration and failed to address complex multimodal correlations effectively. While models such as [26], [27] provided strong feature extraction capabilities, they lacked the dynamic cross-modal interactions that our rotational attention mechanism enables, which allows richer text-image relationships.

Recent models aimed to improve noise suppression and feature extraction but faced similar limitations. For instance, [28], [29] struggled to generalize across domains and overly focused on image credibility. Adversarial networks and ensembling techniques have been explored in prior works such as [30], [31], but these approaches encountered challenges with unstable feature extraction and modality conflicts. Fusion models such as [32], [33] employed complex techniques yet relied on rigid distance metrics, while noise suppression models such as [34], [35] filtered useful signals along with noise. By offering adaptive multimodal fusion and source credibility assessment, our approach significantly enhances fake news detection, particularly in complex scenarios where text and images conflict or come from unreliable sources.

C. Attention-Based Approaches

Attention mechanisms were early adopted in multimodal fake news detection by approaches such as those proposed in [36], [37], combining text, image, and social context features but missing deeper cross-modal relationships. Co-attention and graph networks were explored by work such as [14], [38], [39] to improve text-visual interactions. Similarly, sentiment analysis and entity-centric alignment were integrated by methods such as [40], [41] to capture emotional cues. Despite these advancements, the models remained constrained by rigid structures, limiting their adaptability to dynamic contexts.

Enhanced attention mechanisms, including dual self-attention and ambiguity learning, were introduced by methods such as [42], [12] to improve multimodal integration. Techniques such as self-attention, mutual attention, and multi-head attention were employed by [13], [43]–[45]. Relationship-aware attention, co-attention, and knowledge-augmented features were further advanced by work like [15], [46]–[48].

However, these models often relied on static features, external knowledge, and predefined relationships, which limited their adaptability in rapidly changing and unstructured news environments.

In summary, prior attention models are limited to predefined feature relationships, static knowledge graphs and static attention mechanism. Our model addresses these challenges with dynamic, rotational attention, enabling deeper interactions and flexible relationships, resulting in a more robust system suited for complex, evolving news environments.

III. Method

In this section, we present our proposed multimodal fake new detection framework as illustrated in Figure 2, that leverages both visual and textual features through a novel architecture. It consists of the following key components:

1) **Graph Attention Network (GAT)**: A global GAT models the relationships between news texts and their associated domains. This module leverages:
   - **BERT** [9] embeddings to represent the textual content of news posts.
   - **Word2Vec** [49] embeddings to represent news domains extracted from search results.
2) **Visual Feature Extraction**: Features from images accompanying the news are extracted using the VGG-19 network [50], providing a robust representation of visual content.
3) **Rotational Attention Mechanism**: A unique multi-layer attention mechanism cyclically swaps the roles of query, key, and value across three attention layers. This design enhances the fusion of visual and textual features for more effective detection.
4) **Fake News Classifier**: The integrated outputs are processed by a classifier to predict whether the news is fake or real.

We highlight the key novelty and contributions of this architecture as follows. (1) Novel Use of News Domains: By introducing a global GAT to model the relationships between news domains and their textual content, the framework captures domain-level dependencies, enhancing interpretability and performance. (2) Rotational Attention Mechanism: The innovative attention design enables dynamic interactions between visual and textual modalities, resulting in improved feature fusion. (3) Multi-modal Integration: The integration of both visual features (from VGG-19) and textual features (from BERT and GAT) enables a holistic approach to detecting fake news. In the following, we explain each component in detail.

A. Rich Textual Feature Representation

In this subsection, we first describe the methodology for extracting domain information from search results based on the keywords of a news article; then, we describe how a Graph Attention Network is adopted to utilize embeddings to represent and model the relationships between news texts and their corresponding domains, enhancing the effectiveness of fake news detection.

1) Search Results Domain Extraction: News domain information related to a news article of interest is obtained by searching the keywords of news text online and identifying the most frequently occurring domain names among the search result URLs. The intuition stems from the observation that the presence of certain domains (e.g., cnn.com) can indicate the credibility of a news text. When the keywords of a news text are input into Google, the resulting URL domains can offer context: credible sources tend to appear for real news, while fake news often lacks well-known domains or includes less reputable ones. For example, real news search results typically link to authoritative domains, whereas fake news tends to feature dubious or insignificant domains. Incorporating these domain information helps the model assess news authenticity by providing a broader context for distinguishing between real and fake news.

Specifically, the news text is represented as a sequence of words $T = \{T_i\}_{i=1}^{t}$. The top K frequently occurring words are extracted and input into the Google Custom Search API to get search result URLs. The top common S search result news domains (e.g. wikipedia.org, quora.com) from the URLs are used for further analysis, representing a vector of $1 \times S$.

2) Graph-Based Contextual Analysis: The Graph Attention Network (GAT; [51]) is utilized to model the relationships between news texts and their associated news domains, represented as a bipartite graph. The graph consists of two distinct types of nodes: news text nodes ($v_i \in \mathcal{V}_A$) and news domain nodes ($v_j \in \mathcal{V}_B$), where edges represent relationships between a news text and its top related domains. The news text nodes (v_i) are initialized with BERT [9] embeddings, $\mathbf{h}_i^{(0)} \in \mathbb{R}^{d_{text}}$, while the news domain nodes (v_j) are initialized with Word2Vec [49] embeddings, $\mathbf{h}_j^{(0)} \in \mathbb{R}^{d_{domain}}$.

The edges, denoted by E_{ij}, connect news text nodes in $\mathcal{V}_A$ with news domain nodes in $\mathcal{V}_B$, capturing their relevance. This bipartite graph structure is reflected in the reformulated GAT equations.

To compute the importance of each neighboring node, the attention score e_{ij} between a news text node v_i and a connected news domain node v_j is defined as:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}_A \mathbf{h}_i^{(l)} \parallel \mathbf{W}_B \mathbf{h}_j^{(l)}\right]\right) \quad (1)$$

where $\mathbf{W}_A \in \mathbb{R}^{d \times d_{text}}$ and $\mathbf{W}_B \in \mathbb{R}^{d \times d_{domain}}$ are learnable weight matrices specific to the two node types, $\mathbf{a} \in \mathbb{R}^{2d}$ is a learnable attention vector, and $\parallel$ denotes the concatenation of the transformed features.

The attention scores are normalized using a softmax function to compute the attention coefficients α_{ij}, which determine the contribution of a neighboring node v_j to the feature update of node v_i:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_A(i)} \exp(e_{ik})} \quad (2)$$

where $\mathcal{N}_A(i)$ is the set of neighbors of node v_i in $\mathcal{V}_B$.

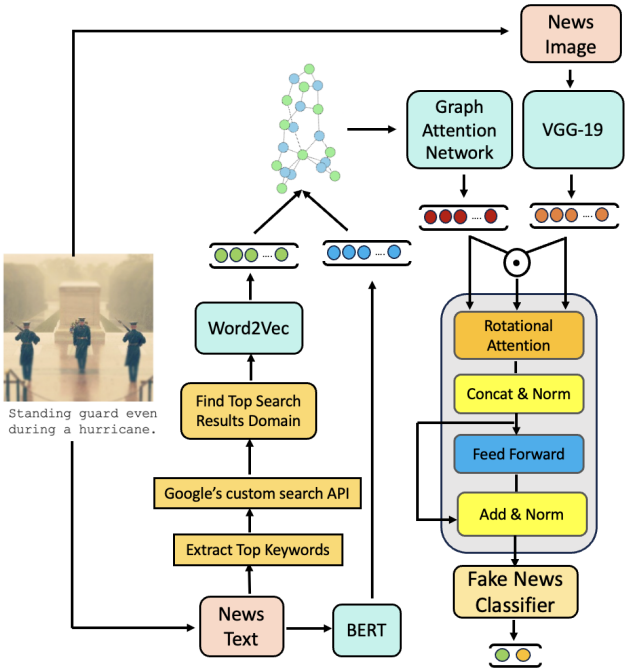

Fig. 2. Architecture of our CAMFeND model. Text features from BERT are enhanced using a Graph Attention Network capturing news post-domain relationships, while visual features come from VGG-19. A rotational attention mechanism exchanges query, key, and value roles between GAT and VGG-19 embeddings. The fused representation undergoes normalization and a feed-forward network before classification into fake or real news. The sample news image is from the Twitter [16] dataset.

The feature of a news text node v_i is updated by aggregating the features of its neighboring news domain nodes $v_j \in \mathcal{V}_B$, weighted by the attention coefficients α_{ij}:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}_A(i)} \alpha_{ij} \mathbf{W}_B \mathbf{h}_j^{(l)}\right) \tag{3}$$

where σ is a non-linear activation function. Similarly, the features of news domain nodes $v_j \in \mathcal{V}_B$ are updated using their neighboring news text nodes $v_i \in \mathcal{V}_A$:

$$\mathbf{h}_j^{(l+1)} = \sigma\left(\sum_{i \in \mathcal{N}_B(j)} \alpha_{ji} \mathbf{W}_A \mathbf{h}_i^{(l)}\right) \tag{4}$$

where $\mathcal{N}_B(j)$ is the set of neighbors of node v_j in $\mathcal{V}_A$.

The GAT is trained using a cross-entropy loss function. After training, the model is frozen, and the learned embeddings of news text nodes ($\mathbf{h}_i$) are used as textual feature representations for subsequent layers in the overall framework.

B. Visual Feature Extraction

For image feature extraction, we use the pre-trained VGG-19 [50] model, a deep convolutional neural network known for its strong performance in image classification tasks. Consisting of 19 layers, with 16 convolutional layers and 3 fully connected layers, it concludes with a softmax layer for classification. To obtain visual features, we add a fully connected layer with ReLU activation after the penultimate layer of VGG-19. This layer generates a $d \times 1$ dimensional VGG-19 feature representation of the input image.

C. The Multimodal Framework

The proposed multimodal framework fuses textual and visual features from news posts using a novel rotational attention mechanism. This section outlines how text and image representations are integrated to form a combined feature vector through a novel rotational attention mechanism.

1) Traditional Attention Mechanism: The standard multi-head self-attention (MSA) [52] block shown in Figure 3(a) uses multi-headed self-attention functions to compute similarity between $d \times 1$ queries (Q), keys (K), and values (V), determining the attention distribution. Multi-Head Attention is composed of multiple attention layers operating in parallel. For m heads, each head performs the following transformations:

$$\mathrm{A}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right) V \tag{5}$$

where, $Q, K, V \in \mathbb{R}^{d_h \times 1}$ and $d_h = \frac{d}{m}$, with d dimension.

The Multi-Head Attention is calculated as:

$$h_j = \mathrm{A}(QW_j^Q, KW_j^K, VW_j^V) \tag{6}$$

$$\mathrm{MHA}(Q, K, V) = \text{concat}(h_1, \ldots, h_m)W^O \tag{7}$$

where, $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{d \times d_h}$ are the j-th head's projection matrices and $W^O \in \mathbb{R}^{d \times d}$ is the output weight matrix.

The fully connected feed-forward network comprises two linear layers separated by a ReLU activation function.

$$\mathrm{FFN}(x) = \max(0, xW_1)W_2 \tag{8}$$

where $x \in \mathbb{R}^{d \times 1}$ is the input to the FFN, $W_1 \in \mathbb{R}^{d \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d}$ are the weights of the FFN, d_{ff} is the hidden dimension of the FFN.

2) Rotational Attention Mechanism: The rotational attention mechanism in Figure 3(b) involves three distinct parallel attention layers, where the roles of query Q, key K, and value V are rotated between the textual and visual embeddings. Let $\mathbf{T}_{gat}$ denote the textual features obtained from the GAT, and $\mathbf{I}_{vgg}$ denote the visual features extracted from the VGG-19 model.

In traditional multi-head attention, multiple parallel heads are used, each applying its own query, key, and value. This approach can be computationally expensive as it requires several attention calculations in parallel, each with separate parameters for Q, K, and V. Moreover, the fixed assignment of roles (Q, K, V) across heads limits the relationships that can be modeled between textual and visual modalities.

Rotational attention improves on this by using a single attention mechanism and rotating the roles of Q, K, and V across three layers. This captures richer interactions between modalities and reduces computational complexity by using fewer parameters (no multi-heads). By rotating roles, the model explores a wider variety of relationships between textual and visual features that would be missed in a fixed-head

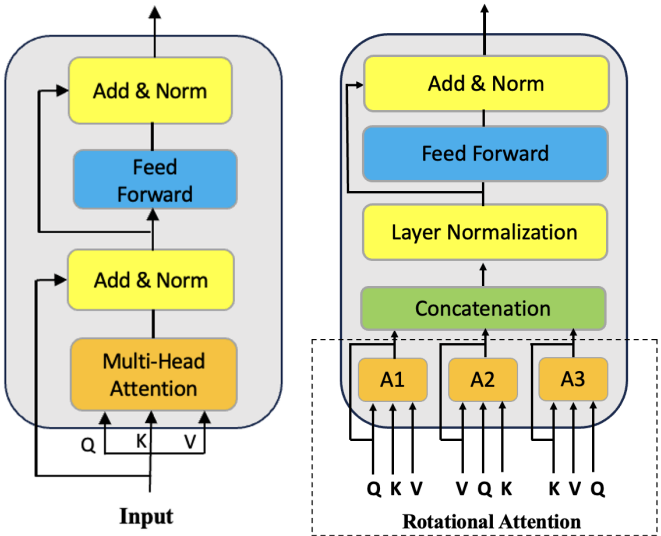

Fig. 3. (a) Self Attention and (b) Rotational Attention: Q, K, and V roles rotate across three attention layers.

approach. The rotational attention mechanism proceeds as follows:

a) Attention 1:

$$\mathbf{A}_1 = \mathrm{A}(\mathbf{I}_{vgg}, \mathbf{T}_{gat}, \mathbf{I}_{vgg} \odot \mathbf{T}_{gat}) + \mathbf{I}_{vgg} \tag{9}$$

In the first attention layer, the query is the VGG-19 embedding $\mathbf{I}_{vgg}$, the key is the GAT embedding $\mathbf{T}_{gat}$, and the value is the element-wise product of the two embeddings, $\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}$.

b) Attention 2:

$$\mathbf{A}_2 = \mathrm{A}(\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}, \mathbf{I}_{vgg}, \mathbf{T}_{gat}) + \mathbf{I}_{vgg} \odot \mathbf{T}_{gat} \tag{10}$$

In the second attention layer, the roles are rotated. The query is the element-wise product $\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}$, the key is the VGG-19 embedding $\mathbf{I}_{vgg}$, and the value is the GAT embedding $\mathbf{T}_{gat}$.

c) Attention 3:

$$\mathbf{A}_3 = \mathrm{A}(\mathbf{T}_{gat}, \mathbf{I}_{vgg} \odot \mathbf{T}_{gat}, \mathbf{I}_{vgg}) + \mathbf{T}_{gat} \tag{11}$$

In the third attention layer, the roles are further rotated. The query is the GAT embedding $\mathbf{T}_{gat}$, the key is the product $\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}$, and the value is the VGG-19 embedding $\mathbf{I}_{vgg}$.

3) Concatenation and Layer Normalization: The outputs from the three attention layers, $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$, are concatenated to form a single vector. This concatenated vector is then passed through a layer normalization process:

$$\mathbf{A}_{concat} = [\mathbf{A}_1; \mathbf{A}_2; \mathbf{A}_3] \tag{12}$$

$$\mathbf{A}_{norm} = \mathrm{LayerNorm}(\mathbf{A}_{concat}) \tag{13}$$

4) Feed-forward Layer and Add & Norm: The normalized vector is processed through a feed-forward layer, followed by an additional add & norm layer to further stabilize the learning process.

$$\mathbf{A}_{ff} = \mathrm{FFN}(\mathbf{A}_{norm}) \tag{14}$$

$$\mathbf{A}_{final} = \mathrm{LayerNorm}(\mathbf{A}_{ff} + \mathbf{A}_{norm}) \tag{15}$$

5) Final Output: The final output $\mathbf{A}_{final}$ from this multimodal framework is used as the combined textual-visual feature representation, which is passed to the fake news classifier for prediction.

D. Fake News Classifier

The combined multimodal representation, becomes the input to the fake news classifier to determine whether a news article is real or fake. It incorporates a fully connected layer with ReLU activation. The predicted probabilities for the k-th post are given by:

$$\hat{y}_k = \sigma(\max(0, W_c \cdot \mathbf{A}_{finalK})W_s) \tag{16}$$

where, $\sigma(.)$ is the softmax function, $\hat{y}_k$ denotes the predicted probabilities, and A_{finalK} is the feature representation of the k-th post. W_c is the fully connected layer parameter and W_s is the softmax layer parameter. We use cross-entropy to calculate the detection loss:

$$\mathcal{L}(\Theta) = -\sum_{k=1}^{N}[Y_k \log(\hat{y}_k) + (1 - Y_k) \log(1 - \hat{y}_k)] \tag{17}$$

where Y_k represents the ground-truth labels of the k-th post and N is the number of posts.

IV. Evaluations

A. Dataset

We evaluate our model CAMFeND on two widely used benchmark datasets in the fake news detection literature: Pheme [1] and Twitter [16]. Pheme contains rumors and non-rumors from five major events, with text, images, and labels. The Twitter dataset includes tweets with text, images, and social context. Given our emphasis on text and image content, we exclude tweets with videos or missing text and images. Pheme is split 80/20 for training/testing, while Twitter provides a pre-split development and test set. These datasets offer a rich environment for evaluating our model with labeled text-image pairs. Table I shows the dataset statistics.

TABLE I
DATA STATISTICS FOR TWO REAL-WORLD DATASETS.

News	Twitter	Pheme
# of Fake News	7898	1972
# of Real News	6026	3830
# of Images	514	3670

B. Implementation Details

Our CAMFeND model is implemented using PyTorch [53], [54], with a model dimension d of 128. We use $K = 20$ for top keywords, $S = 5$ for top news website domains, $m = 1$ for d_{h}, and d_{ff} = 512. Pre-trained BERT [9] and VGG-19 [50] models with frozen parameters are used.

The GAT component includes two hidden layers of dimension 128, optimized using the Adam optimizer [55] with a learning rate of 0.001 and a dropout rate of 0.6. It is trained for

150 epochs with a mini-batch size of 32, and the embeddings are frozen during overall model training.

For model training, we use three hidden layers of dimension 64 for fully connected layers associated with GAT, VGG-19, and rotational attention block embeddings. Our proposed CAMFeND model is trained for 150 epochs with a learning rate of 0.0007, a dropout rate of 0.4, and a mini-batch size of 32 using the Adam optimizer [55]. We use Optuna [56] for hyperparameter tuning with accuracy as the selection criterion.

TABLE II
PERFORMANCE COMPARISON ACROSS TWITTER DATASET

Methods	Acc	Pre	Rec	F1
EANN	0.648	0.709	0.615	0.659
att_RNN	0.664	0.692	0.667	0.679
MVAE	0.745	0.751	0.745	0.748
SpotFake	0.771	0.773	0.773	0.773
SAFE	0.766	0.765	0.764	0.764
SpotFake+	0.790	0.790	0.789	0.789
MCAN	0.809	0.828	0.810	0.819
CAFE	0.806	0.804	0.808	0.806
BMR	0.851	0.885	0.819	0.851
MPL	0.841	0.822	0.860	0.841
CAMFeND	**0.861**	0.898	0.872	**0.885**

TABLE III
PERFORMANCE COMPARISON ACROSS PHEME DATASET

Methods	Acc	Pre	Rec	F1
EANN	0.681	0.696	0.725	0.710
att_RNN	0.850	0.851	0.855	0.853
MVAE	0.852	0.852	0.859	0.855
SpotFake	0.823	0.868	0.863	0.865
SAFE	0.811	0.812	0.828	0.820
SpotFake+	0.800	0.802	0.810	0.806
MCAN	0.865	0.859	0.859	0.859
CAFE	0.861	0.857	0.838	0.847
BMR	0.859	0.824	0.814	0.819
AKA-Fake	0.858	0.918	0.877	0.897
CAMFeND	**0.882**	0.913	0.908	**0.903**

C. Baselines and Results

We evaluate CAMFeND against strong baselines to highlight its effectiveness in fake news detection.

- **EANN** [24]: Derives event-invariant features using a multimodal feature extractor and fake news detector.
- **MVAE** [10]: Uses a variational autoencoder for text and image data with an encoder-decoder structure and a binary classifier to detect fake news.
- **att_RNN** [36]: Embeds attention in a Recurrent Neural Network for the integration of multimodal features.
- **SpotFake** [11]: Employs advanced models such as BERT for textual analysis and VGG-19 for image processing.
- **SAFE** [25]: Uses a similarity-aware multimodal approach to analyze text and visuals.
- **SpotFake+** [26]: Extends SpotFake with a pre-trained XLNet model for textual feature extraction.
- **MCAN** [14]: Dynamically fuses text and image features using a co-attention mechanism.
- **CAFE** [12]: Addresses cross-modal inconsistencies by learning discriminative features through ambiguity learning.
- **BMR** [43]: Uses multi-view feature extraction and an improved Multi-gate Mixture-of-Expert (iMMoE) network for cross-modal learning and fake news detection.
- **MPL** [57]: A multi-modal prompt learning framework for early fake news detection, using pre-trained models and adaptive prompts to generate semantic context rapidly.
- **AKA-Fake** [58]: Utilizes an adaptive knowledge subgraph with reinforcement learning to capture task-relevant knowledge and cross-modal correlations.

Table II and III shows the experimental results of various baseline approaches compared to our CAMFeND model. Early multimodal models like EANN performs slightly better on Pheme compared to Twitter, but it struggles with feature fusion, making it less competitive than models with more advanced multimodal integration methods. Across both datasets, att_RNN performs better than EANN due to its use of attention mechanisms. However, MVAE outperforms both EANN and att_RNN by leveraging a variational autoencoder for more effective multimodal fusion, though it still lags behind models with advanced attention mechanisms.

SpotFake and SpotFake+ leverage pre-trained models like BERT and VGG-19, showing strong results across both datasets. While effective in combining textual and visual features, they are outpaced by more recent models that incorporate attention mechanisms and credibility verification. SAFE uses cross-modal similarity, performing well, but struggles with capturing nuanced interactions, making it less competitive than models with deeper attention mechanisms.

MCAN, with its co-attention mechanism, performs exceptionally well in both datasets, allowing for deep multimodal integration and improving its ability to detect fake news in complex scenarios. CAFE also shows strong performance, particularly on Pheme, though it is slightly less competitive on Twitter. Its cross-modal ambiguity learning helps handle uncertain or ambiguous information. BMR demonstrates effective multimodal fusion, though its performance suggests it could be outperformed by models with more advanced attention mechanisms. MPL and AKA-Fake are among the top performers. MPL leverages multimodal attention, while AKA-Fake benefits from integrating knowledge graphs, both demonstrating solid generalization across datasets, with MPL performing well on Twitter and AKA-Fake excelling on Pheme.

Notably, our proposed CAMFeND model consistently outperforms baseline models on both datasets, highlighting the effectiveness of rotational attention and news domain information in enhancing feature fusion and domain credibility, giving CAMFeND a competitive edge.

D. Ablation Results and Discussions

Table IV presents the ablation study results, analyzing the contribution of key CAMFeND components, particularly rotational attention and news domain information. Both components show a significant impact on performance across the Twitter and Pheme datasets.

TABLE IV
PERFORMANCE OF CAMFEND VARIANTS.

Components	Twitter		Pheme	
	Acc	F1	Acc	F1
CAMFeND¬r	0.782	0.815	0.801	0.835
CAMFeND¬r+sh	0.813	0.838	0.841	0.863
CAMFeND¬r+mh	0.832	0.866	0.850	0.878
CAMFeND¬v	0.743	0.798	0.784	0.817
CAMFeND¬t	0.724	0.767	0.762	0.792
CAMFeND¬n	0.803	0.821	0.827	0.846
CAMFeND	**0.861**	**0.885**	**0.882**	**0.903**

1) Impact of Rotational Attention: Removing the rotational attention mechanism (CAMFeND¬r) results in a significant drop in performance across both datasets, with Twitter showing an accuracy drop and Pheme experiencing a similar decline. This indicates that rotational attention plays a crucial role in enabling dynamic cross-modal interactions between text and images.

Using a single transformer unit, both single-head attention (CAMFeND¬r+sh) and multi-head attention (CAMFeND¬r+mh) improve over the model without rotational attention. In both Twitter and Pheme datasets, these variants boost accuracy but still fall short of the complete model (CAMFeND), which achieves higher accuracy in both datasets.

While multi-head attention offers advantages over single-head attention, it lacks the dynamic nature of rotational attention, which enables diverse interactions between the query, key, and value components. The rotational attention mechanism in CAMFeND enhances the model's ability to explore rotational interaction of input modalities, leading to deeper interactions and better understanding of cross-modal signals, resulting in higher accuracy and performance across both datasets.

2) Effect of Component Removal: Removing the visual component (CAMFeND¬v) or the textual component (CAMFeND¬t) leads to significant drops in performance for both datasets. On Twitter, removing the visual component causes a notable drop in accuracy, while removing the textual component similarly impacts performance. On Pheme, removing either component shows a similar trend, confirming that both modalities provide essential information for accurate detection in multimodal fake news detection.

3) Role of News Domains: The inclusion of news domain information proves to be a critical factor in improving the model's robustness. When news domains are omitted (CAMFeND¬n), the model relies solely on BERT embeddings for textual features, leading to a drop in performance in both datasets. This shows that news domain information adds a crucial layer of source reliability assessment, helping the model filter out unreliable sources and reducing false detections that may arise when relying purely on content.

V. CONCLUSIONS

We presented CAMFeND, a novel multimodal fake news detection model that combines rotational attention and news domain information. By rotating the roles of query, key, and value between text and image features, our model captures deeper cross-modal interactions for more accurate detection. The integration of news domain information enhances robustness by providing broader contextual cues from associated domains. Comprehensive evaluations on the Twitter and Pheme datasets show that CAMFeND consistently outperforms baseline models.

ACKNOWLEDGEMENT

This material is based upon work supported in part by the Department of Energy under Award Number DE-CR0000003.

REFERENCES

[1] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*. Springer, 2017, pp. 109–123.

[2] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[3] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 1103–1108.

[4] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1867–1870.

[5] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1751–1754.

[6] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan *et al.*, "A convolutional approach for misinformation identification." in *IJCAI*, 2017, pp. 3901–3907.

[7] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," 2016.

[8] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional lstm-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2019.

[9] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[10] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915–2921.

[11] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 2019, pp. 39–47.

[12] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM web conference 2022*, 2022, pp. 2897–2905.

[13] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information processing & management*, vol. 60, no. 1, p. 103120, 2023.

[14] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569.

[15] H. Yang, J. Zhang, L. Zhang, X. Cheng, and Z. Hu, "Mran: Multimodal relationship-aware attention network for fake news detection," *Computer Standards & Interfaces*, vol. 89, p. 103822, 2024.

[16] C. Boididou, S. Papadopoulos, D. T. Dang Nguyen, G. Boato, M. Riegler, A. Petlund, and I. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," 10 2016.

[17] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *2012 IEEE symposium on security and privacy*. IEEE, 2012, pp. 461–475.

[18] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 651–662.

[19] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan *et al.*, "A convolutional approach for misinformation identification." in *IJCAI*, 2017, pp. 3901–3907.

[20] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 518–527.

[21] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019.

[22] M. Cheng, S. Nazarian, and P. Bogdan, "Vroc: Variational autoencoder-aided multi-task rumor classifier based on text," in *Proceedings of the web conference 2020*, 2020, pp. 2892–2898.

[23] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 549–556.

[24] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "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*, 2018, pp. 849–857.

[25] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, 2020.

[26] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 10, 2020, pp. 13 915–13 916.

[27] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 540–547.

[28] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 557–565.

[29] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21 503–21 517, 2022.

[30] P. Wei, F. Wu, Y. Sun, H. Zhou, and X.-Y. Jing, "Modality and event adversarial networks for multi-modal fake news detection," *IEEE Signal Processing Letters*, vol. 29, pp. 1382–1386, 2022.

[31] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "Semi-fnd: Stacked ensemble based multimodal inferencing framework for faster fake news detection," *Expert systems with applications*, vol. 215, p. 119302, 2023.

[32] Z. Zeng, M. Wu, G. Li, X. Li, Z. Huang, and Y. Sha, "An explainable multi-view semantic fusion model for multimodal fake news detection," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1235–1240.

[33] J. Wang, J. Zheng, S. Yao, R. Wang, and H. Du, "Tlfnd: A multimodal fusion model based on three-level feature matching distance for fake news detection," *Entropy*, vol. 25, no. 11, p. 1533, 2023.

[34] Y. Gu, I. Castro, and G. Tyson, "Detecting multimodal fake news with gated variational autoencoder," in *Proceedings of the 16th ACM Web Science Conference*, 2024, pp. 129–138.

[35] S. Zhong, S. Peng, X. Liu, L. Zhu, X. Xu, and T. Li, "Ecarnet: enhanced clue-ambiguity reasoning network for multimodal fake news detection," *Multimedia Systems*, vol. 30, no. 1, p. 55, 2024.

[36] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.

[37] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[38] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, p. 102437, 2021.

[39] S. Qian, J. Hu, Q. Fang, and C. Xu, "Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 3, pp. 1–23, 2021.

[40] Q. Jing, D. Yao, X. Fan, B. Wang, H. Tan, X. Bu, and J. Bi, "Transfake: multi-task transformer for multimodal enhanced fake news detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[41] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, and G. Xu, "Entity-oriented multi-modal alignment and fusion network for fake news detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 3455–3468, 2021.

[42] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, and Y. Zhang, "Mfan: Multi-modal feature-enhanced attention networks for rumor detection." in *IJCAI*, vol. 2022, 2022, pp. 2413–2419.

[43] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge, "Bootstrapping multi-view representations for fake news detection," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 5384–5392.

[44] Y. Guo, "A mutual attention based multimodal fusion for fake news detection on social network," *Applied Intelligence*, vol. 53, no. 12, pp. 15 311–15 320, 2023.

[45] L. Wu, Y. Long, C. Gao, Z. Wang, and Y. Zhang, "Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection," *Information Fusion*, vol. 100, p. 101944, 2023.

[46] X. Liu, P. P. Li, H. Huang, Z. Li, X. Cui, W. Deng, Z. He *et al.*, "Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms," in *ACM Multimedia 2024*, 2024.

[47] Z. Yi, S. Lu, X. Tang, J. Wu, and J. Zhu, "Maccn: Multi-modal adaptive co-attention fusion contrastive learning networks for fake news detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6045–6049.

[48] C. Yin and Y. Chen, "Multi-modal co-attention capsule network for fake news detection," *Optical Memory and Neural Networks*, vol. 33, no. 1, pp. 13–27, 2024.

[49] K. W. Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[52] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[53] J. Hu, S. Qian, Q. Fang, Y. Wang, Q. Zhao, H. Zhang, and C. Xu, "Efficient graph deep learning in tensorflow with tf_geometric," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3775–3778.

[54] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[56] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[57] W. Hu, Y. Wang, Y. Jia, Q. Liao, and B. Zhou, "A multi-modal prompt learning framework for early detection of fake news," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 651–662.

[58] L. Zhang, X. Zhang, Z. Zhou, F. Huang, and C. Li, "Reinforced adaptive knowledge learning for multimodal fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 777–16 785.