GAMED: Knowledge Adaptive Multi-Experts Decoupling for Multimodal Fake News Detection

Lingzhi Shen
University of Southampton
Southampton, United Kingdom
l.shen@soton.ac.uk

Yunfei Long
University of Essex
Essex, United Kingdom
yl20051@essex.ac.uk

Xiaohao Cai
University of Southampton
Southampton, United Kingdom
x.cai@soton.ac.uk

Imran Razzak
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
The University of New South Wales
Sydney, Australia
imran.razzak@unsw.edu.au

Guanming Chen
Kang Liu
University of Southampton
Southampton, United Kingdom
gc3n21@soton.ac.uk
kl2y21@soton.ac.uk

Shoaib Jameel*
University of Southampton
Southampton, United Kingdom
M.S.Jameel@southampton.ac.uk

Abstract

Multimodal fake news detection often involves modelling heterogeneous data sources, such as vision and language. Existing detection methods typically rely on fusion effectiveness and cross-modal consistency to model the content, complicating understanding how each modality affects prediction accuracy. Additionally, these methods are primarily based on static feature modelling, making it difficult to adapt to the dynamic changes and relationships between different data modalities. This paper develops a significantly novel approach, GAMED, for multimodal modelling, which focuses on generating distinctive and discriminative features through modal decoupling to enhance cross-modal synergies, thereby optimizing overall performance in the detection process. GAMED leverages multiple parallel expert networks to refine features and pre-embed semantic knowledge to improve the experts' ability in information selection and viewpoint sharing. Subsequently, the feature distribution of each modality is adaptively adjusted based on the respective experts' opinions. GAMED also introduces a novel classification technique to dynamically manage contributions from different modalities, while improving the explainability of decisions. Experimental results on the Fakeddit and Yang datasets demonstrate that GAMED performs better than recently developed state-of-the-art models. The source code can be accessed at https://github.com/slz0925/GAMED.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Computer vision**.

*Shoaib Jameel is the corresponding author.

WSDM '25, March 10–14, 2025, Hannover, Germany

ACM ISBN 979-8-4007-1329-3/25/03
https://doi.org/10.1145/3701551.3703541

Keywords

Fake News Detection; Multimodal Learning; Pattern Recognition; Mixture of Experts; Explainable AI

ACM Reference Format:
Lingzhi Shen, Yunfei Long, Xiaohao Cai, Imran Razzak, Guanming Chen, Kang Liu, and Shoaib Jameel. 2025. GAMED: Knowledge Adaptive Multi-Experts Decoupling for Multimodal Fake News Detection. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701551.3703541

1 Introduction

Imagine encountering a social media post with a seemingly innocuous image and a captivating new headline. Unfortunately, a sinister truth lurks beneath this alluring facade – it is well-crafted fake news [16, 22]. Nowadays everyone is an editor and everyone can publish news – especially on social media [50]. As a result, there is an escalating threat of multimodal fake news [29, 64], a potent weapon that weaponizes the synergy of text and visuals to manipulate public discourse and erode trust in information [60, 62]. According to [74], fake news is defined as "purposefully crafted, sensational, emotionally charged, misleading or fabricated information that mimics the form of mainstream news".

Traditional multimodal detection [4], typically rely on basic fusion techniques [43, 67], striving to decipher the complex interactions within multimodal narratives [40]. The result is that they fail to capture the nuances that distinguish genuine news from its fabricated counterparts [10, 59]. This critical gap in detection capabilities stems from several inherent limitations, for instance, many current fusion techniques suffer from feature suppression issues [70], hindering the model's ability to grasp the intricate dance between textual and visual elements within a single content. Another reason is the over-reliance on the inherent capabilities of pre-trained models without any additional refinement step for feature representations in the pipeline [3, 47], which is prone to lose information that is crucial for classification. Similarly, existing methods often focus solely on identifying consistency between modalities [55], and they fail to consider utilizing discriminative features as a complement. However, in the real world, many fabricators

have learned to bypass the detection of previous consistency-centric models [2, 46]. Moreover, the black-box nature of decision-making processes in existing models shrouds the perspectives and contributions of each modality in mystery. This lack of transparency hinders interpretability [18, 52] and erodes user trust [49].

Our primary objective is to create a novel approach called GAMED, which autonomously models fake news within a multimodal context, enhancing the current state-of-the-art by overcoming some of its intrinsic limitations. In pursuit of this goal, we utilize publicly accessible multimodal datasets. For example, [69] presented the BMR model for multimodal fake news detection. While BMR promotes in-depth multimodal analysis and improved feature extraction, it suffers from some key shortcomings. Firstly, BMR faces challenges in integrating multimodal data and dynamically weighting different modalities. Additionally, the bootstrapping process for multi-view representations hinders BMR's interpretability. Most importantly, BMR fails to invoke valuable real-world factual knowledge as a reference when the model encounters confusion [24, 44]. Another recent work by Xuan et al. [65] investigated multimodal fake news detection with their LEMMA system which combines vision language models with external knowledge, but it lacks the feature refinement strategy, and this may also introduce additional noise.

Technical contributions: GAMED, based on modal-decoupling modelling, exploits the potential of cross-modal synergies to improve detection performance, which is distinct from existing multimodal fake news detection methods based on consistency learning or fusion-only strategies, such as those in [20, 33, 54]. GAMED combines the characteristics of expert networks and AdaIN [26] to perform progressive feature refinement to obtain more discriminative and distinctive feature representations, providing a novel paradigm for dynamic screening and optimization of multimodal data. GAMED demonstrates that external knowledge, e.g., semantic knowledge graph information encoded in pre-trained language models [13] is beneficial to help models better understand the complex relationships and contexts in fake news, and extends with the influence from text to other modalities. GAMED introduces a novel decision-making method, which is conducive to improving the transparency and explainability of the decision-making process. GAMED attains better detection performance on the publicly available Fakeddit[1] and Yang datasets [68], presenting a novel solution for automated fake news detection.

2 Related Work

Unimodal fake news methods: Unimodal fake news detection [8, 15, 51] has made significant strides in tackling fake news [53]. Traditional machine learning algorithms, such as SVM, Decision Trees, and Naïve Bayes [6, 21, 27], along with modern deep learning approaches, including CNN, RNN, and LSTM [23, 37], have been extensively compared and studied. For text analysis, transformer-based models such as BERT [12] have paved the way for advancements such as RoBERTa and GPT-3 [5], enhancing the detection of subtle linguistic cues. Similarly, sophisticated architectures such as ResNet and ViTs [28] have revolutionized image analysis, enabling the identification of manipulated visuals. However, unimodal approaches have an inherent weakness: they struggle against multimodal fake news that blends text, images, and other media for a more convincing narrative [73].

Multimodal fake news methods: Multimodal fake news detection has emerged as a critical research area, which analyzes data from multiple sources – text, images, videos, and social context – to form a more holistic view of the information [25]. By leveraging the complementary strengths, these approaches aim to uncover discrepancies that might not be evident when analyzing text alone. This field has gravitated towards leveraging the synergistic potential of advanced fusion techniques and models such as VisualBERT [32], ViLBERT [17], and LXMERT [58], which facilitate dynamic, context-aware integration of text and images through self-attention mechanisms. These models have significantly advanced the capacity to understand and analyze the complex interplay between modalities, often focusing on exploiting cross-modal dynamics and consistencies as potent indicators of misinformation. Moreover, the integration of external knowledge sources [14, 39], through methods like knowledge graph embeddings [16], has provided additional context for verifying claims, enhancing the models' ability to discern truth from deception. Despite these advancements, multi-modal detection still faces notable challenges, particularly in processing effectiveness and the adaptive generalization to new forms of fake content, such as deepfakes. The quest for explainability [71] in these complex models remains an ongoing challenge. Moreover, existing multimodal research predominantly focuses on innovative fusion techniques, while how to leverage the distinctive potential of each modality remains an unresolved issue [38, 41].

3 Our Novel GAMED Model

As depicted in Figure 1, GAMED is a novel modality-decoupling design for detecting fake news across textual and visual modalities. The process starts with extracting features from text and images, followed by a stage that simulates expert review and opinions using the MMoE-Pro network to refine feature representations. Subsequently, the distribution adjustment stage, guided by the AdaIN adaptive mechanism, dynamically fine-tunes the impact of each modality, giving precedence to the most pertinent and trustworthy data. Finally, a novel voting system with veto power is introduced in the decision-making stage by combining consensus-based and confidence-based evaluation methods. The entire GAMED workflow also benefits from the semantic information encoded (KE) by pre-trained language models that encode structured (e.g., knowledge graphs) and unstructured text information. In Algorithm 1, we meanwhile present the detailed pseudo-code of GAMED.

Feature Extraction: We represent our multimodal news data as a collection $\mathcal{N} = [\mathbf{I}, \mathbf{T}] \in \mathcal{D}$, where $\mathbf{I}, \mathbf{T}, \mathcal{D}$ are the image, the text, and the dataset, respectively. Each data point within $\mathcal{D}$ allows us to analyze the interplay between visual content and textual narrative. We exploit the Inception-ResNet-v2 (IRNv2) [57] as a feature extractor to extract image patterns (IP), denoted by f_{ip}. We add a special filter BayarConv [7] as an early layer. Our intuition is that when images are tampered with, they often leave subtle traces of forgery that are not easily detected by traditional convolutions, such as artefacts, lighting, and texture. We capture image semantic (IS) features on the global and local details of the image by combining

[1] https://github.com/entitize/Fakeddit

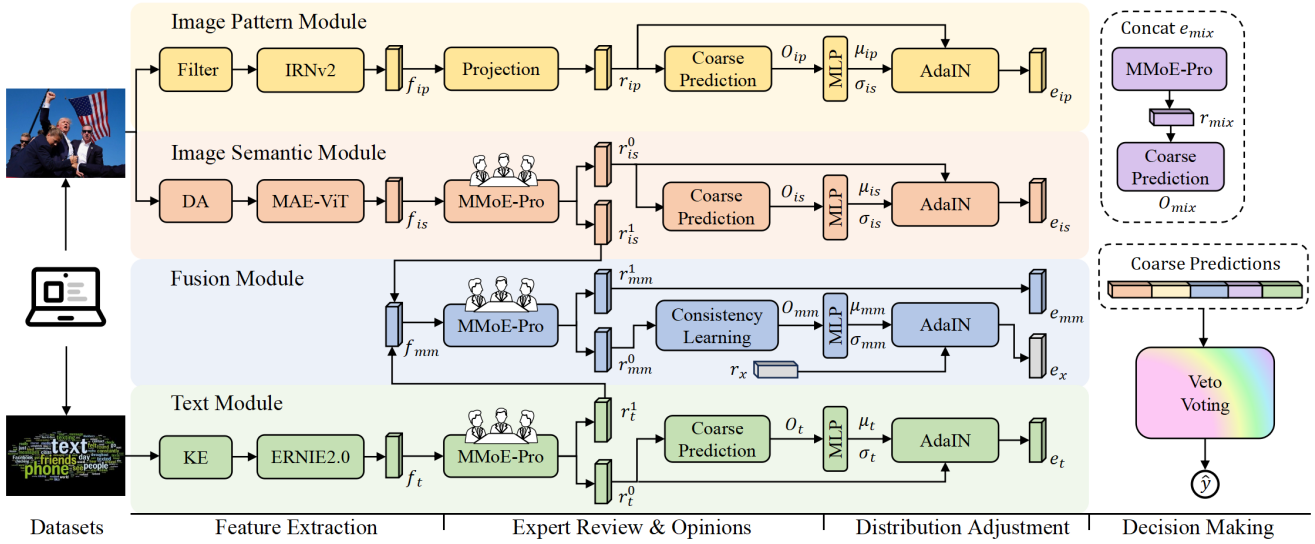

Figure 1: Starting from raw data, the GAMED's modality-specific pipeline performs feature extraction and progressive refinement. The knowledge enhancement mechanism provides an external background to the architecture. During the expert review stage, features are selected and coarse predictions are made. The AdaIN component then adaptively adjusts the feature distribution. The decisive voting stage orchestrates the final classification.

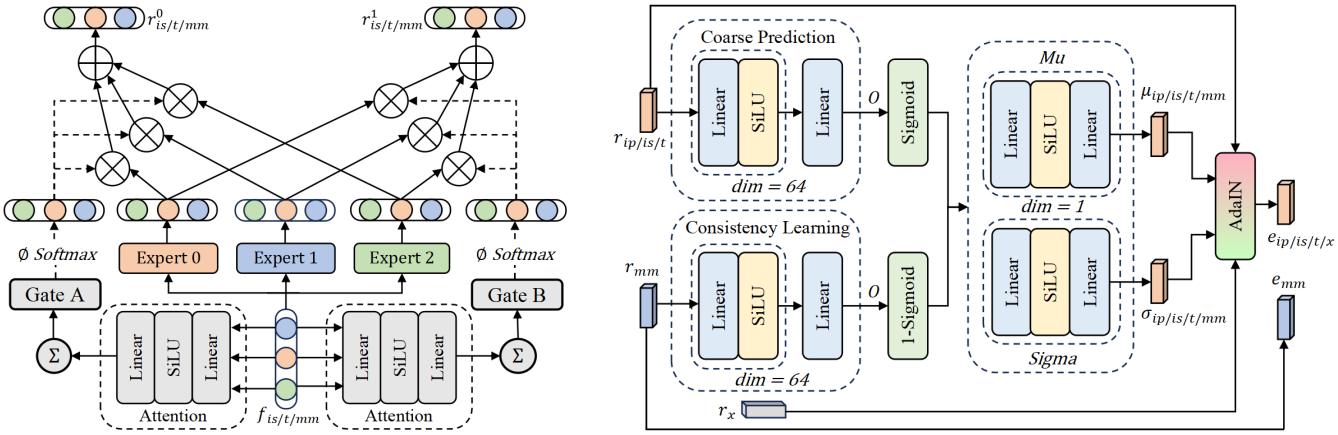

Figure 2: Left: The configuration of MMoE-Pro and the flow of processing representations. Right: The pipeline of four modules from coarse prediction to adaptive feature distribution adjustment to obtain enhanced representations.

ViT with a masked auto-encoder model (MAE) [19, 36], denoted as f_{is}. The use of data augmentation (DA), such as rotation, flipping, and scaling, enables the model to better generalize inconsistencies often encountered in fake images to enhance the robustness and diversity of image semantic data.

We exploit the ERNIE2.0 [56] model to extract the text (T) representations, denoted as f_t. ERNIE2.0 includes several advantages such as modelling sentence-level relations (in addition to word-level), and large-scale semantic knowledge stored in knowledge graphs. With these improvements, ERNIE2.0 can better evaluate the consistency of text content with visual information in images, and analyze facts and entity relationships in text. The structured knowledge encoded in ERNIE2.0 helps enhance the reasoning ability of the entire architecture in global synergies across modalities.

Expert Review and Opinions: The workflow of the novel expert network is depicted in the model in Figure 2 (left). In this stage, we simulate the scenario of experts with rich expertise to review. The expert networks of different modalities accept features provided by the extractors of their respective modules, and then the mixture of experts will jointly review and select these features, and provide preliminary predictions.

Our MMoE-Pro upgrades the traditional MMoE [45] mainly by introducing token attention and relaxing the softmax constraint in its gating mechanism. In particular, suppose the input f, which

can be from IS, T, or fusion (MM) modules, consists of multiple tokens, the process begins by calculating the importance score α_i for each token representation using a shared-weight MLP A, formalized as $\alpha_i = A(\text{token}_i)$. The normalized importance score β_i is calculated as $\beta_i = \frac{\alpha_i}{\sum_j \alpha_j}$. This normalization allows for the aggregation of token representations into a unified form, where the aggregated representation $\tilde{f}$ is computed as a weighted sum of the token representations: $\tilde{f} = \sum_i \beta_i \cdot \text{token}_i$. This approach enhances the model's ability to identify key features by dynamically evaluating the importance of different input features, thereby optimizing the feature-sharing process in multi-task learning.

Furthermore, we adjust the gating mechanism by lifting the strict positivity and normalization constraints traditionally imposed by softmax. Specifically, the new weights $w_{t,i}(f)$, used for determining the contribution of each expert i for task t, are derived directly from the raw scores, allowing weights to take on negative values or exceed one. It is because there is no necessity to use constraints to force the expert's contribution to be positive in our setting. The final output of the MMoE-Pro model is now formalized as

$$\text{MMoE-Pro}_t(f) = \sum_{i=1}^{N} w_{t,i}(\tilde{f}) E_i(f). \tag{1}$$

Here, $E_i(f)$ represents the output of the i-th expert. Ultimately, the output features $r = \text{MMoE-Pro}_t(f)$ serve as the feature representation jointly selected by the expert team, including $\left[r_{is}^0, r_{is}^1\right]$, $\left[r_t^0, r_t^1\right]$ and $\left[r_{mm}^0, r_{mm}^1\right]$. Here, r_{is}^1 and r_t^1 serve as the initial features for the fusion module, f_{mm}, and as inputs to the expert network in this module. These new improvements can more flexibly and effectively enhance the ability of the expert network to dynamically allocate computing resources across different tasks and modalities.

We use an MLP classifier to accept the refined representations from the MMoE-Pro and Projection (f_{ip}) to achieve the coarse prediction function. This is depicted in the steps of coarse prediction and consistency learning in Figure 2. For each refined representation r from the expert network, we perform a 64-dimensional feature reduction and produce the classification output. This combined process can be expressed as $O = \text{MLP}(r)$, where O represents the coarse prediction output.

Distribution Adjustment: We depict this stage in Figure 2 (right). The coarse prediction results from the previous stage are calculated as mean and standard deviation as an acceptable input form for AdaIN. AdaIN then adaptively adjusts the feature distribution according to the contribution of each modality. This step ensures that the most relevant and reliable information is prioritized.

We first calculate the parameters, mean μ and standard deviation σ, required by AdaIN. Unlike the standard AdaIN approach, our μ and σ are generated through MLP networks rather than being directly extracted from the style features. Specifically, for each output O from the coarse prediction, we use MLPs to generate the mean and standard deviation, denoted as $\mu = \text{MLP}_\mu(\text{sigmoid}(O))$ and $\sigma = \text{MLP}_\sigma(\text{sigmoid}(O))$, where MLP_μ and MLP_σ represent the MLPs for calculating the mean and standard deviation, respectively. The adjustment process of AdaIN is then formalized as

$$e = \sigma(r - \mu_r)/\sigma_r + \mu. \tag{2}$$

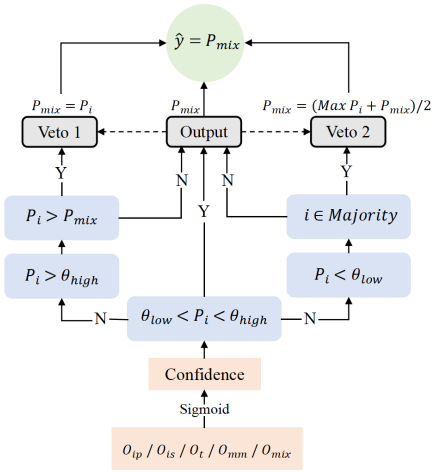

Figure 3: Our novel veto model.

Here, μ_r and σ_r are respectively the mean and standard deviation of the input feature r, and e is the adjusted feature. The process of combining information from each modality in AdaIN can be simplified as $e = \text{AdaIN}(r, \mu, \sigma)$. At this stage, for the mean and standard deviation calculation of the prediction output from consistency learning, we use $(1 - \text{sigmoid}(O))$ to invert it so that adjust the distribution of irrelevant representation r_x to e_x. Our intuition is that certain irrelevant information, such as emotional language, lengthy background introduction, complex rhetoric, etc., can also serve as clues for detecting fake news. Therefore, we decouple relevance from consistency learning and leverage irrelevance as a synergistic supplement. We directly use the fusion representation r_{mm} as the adjusted representation e_{mm} of the fusion module.

Veto Voting: The novel veto classifier is depicted in Figure 3. Before the voting mechanism is triggered, the enhanced representations of all modalities produced by AdaIN are concatenated and denoted as e_{mix} (Figure 1), and then the MMoE-Pro performs the same refinement process to obtain r_{mix} (Figure 1) and O_{mix} successively. Finally, this concatenated prediction output and the previous coarse prediction output of each modality are used as input into the voting stage. Our novel veto voting combines the inspiration of thresholds and confidence to dynamically manage the contributions and conflicts of each modality prediction to ensure the reliability and transparency of the decision-making.

We define two thresholds to distinguish between high confidence and low confidence, where θ_{high} and θ_{low} define the high confidence and low confidence thresholds, respectively, used to determine if a module's prediction can be used as a decision basis. For the prediction output O_i of each module i, we apply the sigmoid function to convert it into a confidence probability value as $P_i = \text{sigmoid}(O_i)$ where P_i is the confidence of the prediction from module i. Let the confidence of the concatenated output be P_{mix}, initially set as $P_{mix} = \text{sigmoid}(O_{mix})$. Suppose P_{mix} is used as the basis for comparing other module confidences and for the final decision. Let O_i and O_{mix} be the raw outputs of the module prediction and the concatenated output, respectively. Let the majority class be the class decided by the majority of module decisions; whether module i belongs to the majority class can be determined

Algorithm 1 PyTorch-style Pseudocode of GAMED

Require: Dataset: $\mathcal{D}$; training epochs: N; batch size: B; and learning rate: η.
Ensure: Model parameters: Θ

```
Define loss function: L_BCE
Define optimizer: AdamW
for epoch in range(N) do
    Load batch data (T, I, y)
    Forward Propagation:
        f = t, is, ip = extract(T, I)
        r = experts(t, is, ip)
        r_mm = experts(mm = t + is)
        O = class(reduc(r_0, r_1, ..., r_n))
        μ, σ = comp(O, 1 − O)
        e = AdaIN(r, μ, σ)
        f_mix = concat(e_0, e_1, ..., e_n)
        r_mix = experts(f_mix)
        O_mix = class(reduc(r_mix))
        ŷ = veto(O_0, O_1, ..., O_n)
    Back Propagation:
        L = ComputeLoss(O_0, O_1, ..., O_n)
        L.backward()
        optimizer.step()
end for
```

by comparing it to other module predictions. We iterate through each module and update the concatenated output according to the following rules.

Rule 1: For the initial concatenated output, we denote the rule as $P_{mix} = \text{sigmoid}(O_{mix})$.

Rule 2: If the module confidence P_i is greater than the high confidence threshold and $P_i > P_{mix}$, then replace the concatenated output with the output from module i. This is denoted as $P_{mix} = P_i$ if $P_i > \theta_{\text{high}}$ and $P_i > P_{mix}$.

Rule 3: If the module confidence P_i is less than the low confidence threshold, and module i belongs to the majority class, then ignore any output in the majority class, and reconsider the maximum output from all modules. This is denoted as $P_{mix} = \frac{1}{2}(P_{mix} + \max P_i)$ if $P_i < \theta_{\text{low}}$ and i belongs to the majority class.

Rule 4: If the confidence is between the thresholds, maintain the concatenated output. This is denoted as $P_{mix} = P_{mix}$ if $\theta_{\text{low}} \leq P_i \leq \theta_{\text{high}}$.

4 Experiments and Results

To rigorously evaluate the efficacy of GAMED in detecting fake news, we conducted extensive experiments. This section details the experimental framework, evaluation criteria, and the notable results obtained. The objective is to determine whether GAMED outperforms recent robust models and to assess the contribution of each component through ablation studies. A qualitative analysis also demonstrates the transparency of the decision-making process.

4.1 Experimental Setup

Datasets: We conducted training, validation, and test of GAMED and other models on two publicly accessible datasets: Fakeddit and Yang. Fakeddit is a vast collection with over one million labelled samples, classified as real or fake news. It offers a balanced division, consisting of 628,501 fake news instances and 527,049 real news instances. Derived from a wide range of 22 subreddits, Fakeddit provides a rich diversity of domains and topics, mirroring the real-world scenario. Fakeddit is a dataset that offers fine-grained categories. For our model, which focuses on binary classification tasks, we utilize only the 2-way labels. The Yang dataset includes 20,015 news articles, with 11,941 marked as fake and 8,074 as real. The fake news is sourced from more than 240 websites, while the genuine news is obtained from reputable, authoritative outlets like the New York Times and Washington Post. The dataset used in Ying et al. [69] is not publicly accessible due to strict API restrictions on obtaining image data from Twitter and Weibo. Our attempts to contact the authors were unsuccessful.

Settings: The Fakeddit dataset comprises 563,612 training samples, 58,798 validation samples, and 59,271 test samples. In contrast, Yang's dataset contains 4,655 training samples, 582 validation samples, and 583 test samples. Each text has a corresponding image. In processing image data, we utilize two pre-trained models: *mae-pretrain-vit-base* for semantic analysis and *pytorch-InceptionResNetV2* for pattern recognition. Text data is processed with *ernie-2.0-base-en*. MAE-ViT and ERNIE have a hidden dimension of 768, with their parameters kept frozen. Our preprocessing steps aim to optimize the handling of input data. This process includes resizing all images to a consistent size of 224×224 pixels. We also set a maximum tokenization length of 197 for both text and image data. All MLPs in GAMED include one hidden layer and SiLU activation function. We use AdamW optimizer with 1×10^{-4} learning rate. The model typically reaches peak accuracy within 9–10 epochs on Fakeddit and 5–7 epochs on Yang.

Evaluation Metrics: To align with comparative models, we use the widely accepted metrics: Accuracy (Acc), Recall (R), Precision (P), and F1 Score for this task.

Comparative Models: The selection of baselines is shown in Table 1. For the Fakeddit dataset, our model is benchmarked against recent robust comparative models. The EANN model [62] combines Text-CNN and VGG-19 to filter event-specific features while retaining event-independent features, enhancing adaptability to new events. The MVAE [30] model was chosen for its use of a multimodal variational autoencoder that learns joint representations from text and image data, improving detection by capturing the interactions between modalities with LSTM and VGG-19. The BMR model [69], on the other hand, bootstraps multi-view representations to refine and reweigh features from text and image for superior fake news detection performance. The MMBT [31] integrates text features from BERT with image features from ResNet-152 using a single transformer, while MTTV [61] employs a dual-level visual feature extraction approach with BERT and ResNet to bolster the synergy between textual and visual data. The CLIP and LLaVA hybrid architecture [34] employs LoRA-based fine-tuning strategies and knowledge transfer, effectively enhancing multimodal fact verification by integrating visual and textual evidence. ELD-FN [42] combines ViBERT-generated multimodal embeddings with sentiment analysis through an ensemble learning approach.

Dataset	Method	Acc	P	R	F1
Fakeddit	EANN	87.50	90.43	88.11	89.26
	MVAE	88.75	90.11	91.39	90.74
	ELD-FN	88.83	93.54	90.29	91.89
	MMBT	91.11	92.74	92.51	92.63
	MTTV	91.88	93.48	93.03	93.25
	BMR	91.65	**94.34**	92.88	93.61
	CLIP+LLaVA	92.54	93.85	91.24	92.53
	GAMED	**93.93**	93.55	**93.71**	**93.63**
Yang	EANN	85.54	86.37	84.32	85.33
	MVAE	90.85	91.58	90.94	91.26
	SAFE	92.27	93.75	93.53	93.64
	TI-CNN	92.48	92.20	92.77	92.10
	BMR	94.34	95.23	93.59	94.15
	BERT+MVNN	95.68	96.44	95.93	96.18
	MCNN	96.30	97.29	96.44	96.86
	GAMED	**98.46**	**98.31**	**98.59**	**98.43**

Table 1: Comparison between our GAMED and state-of-the-art. The best results (%) are highlighted in bold.

Fakeddit Dataset							
Method	Acc	Fake News			Real News		
		P	R	F1	P	R	F1
LLaVA (Direct)	0.663	0.588	0.797	0.677	0.777	0.558	0.649
LLaVA (CoT)	0.673	0.612	0.400	0.484	0.694	0.843	0.761
GPT-4 (Direct)	0.677	0.598	0.771	0.674	0.776	0.606	0.680
GPT-4 (CoT)	0.691	0.662	0.573	0.614	0.708	0.779	0.742
GPT-4V (Direct)	0.734	0.673	0.723	0.697	0.771	0.742	0.764
GPT-4V (CoT)	0.754	0.858	0.513	0.642	0.720	**0.937**	0.814
FacTool	0.506	0.476	0.834	0.606	0.624	0.232	0.339
InstructBLIP	0.726	0.760	0.489	0.595	0.715	0.892	0.793
LEMMA	0.824	0.835	0.727	0.777	0.818	0.895	0.854
GAMED	**0.939**	**0.954**	**0.944**	**0.949**	**0.917**	0.930	**0.923**

Table 2: Comparison between GAMED and large language models on Fakeddit. The best results are highlighted in bold.

For the Yang dataset, in addition to using the same baselines (i.e., EANN, MVAE, and BMR) as for the Fakeddit dataset, we further introduced several advanced architectures. For instance, the MCNN [66] model integrates text and visual features through five sub-networks, including BERT, BiGRU, ResNet50, ELA algorithms, and attention mechanisms, detecting inconsistencies in multimodal data by measuring similarity and identifying visual tampering. The SAFE [72] model extracts features from news text and visuals using an extended Text-CNN and Image2Sentence model, identifying mismatches between text and images. TI-CNN [68] uses parallel CNNs to extract and fuse explicit and latent features from text and images, establishing itself as an initial benchmark for the Yang dataset. Finally, MVNN [48] combines frequency and pixel domain visual information using CNN and multi-branch CNN-RNN networks, along with attention mechanisms to dynamically integrate physical and semantic features in fake news images.

4.2 Overall Results

The results in Table 1 demonstrate that our GAMED is quantitatively superior in performance when compared with different competitive models including those that are recently developed. On Fakeddit, GAMED achieves 93.93% accuracy, surpassing the state-of-the-art open-source detection scheme MTTV by 2.05% and LoRA-fine-tuned CLIP and LLaVA combination by 1.39%. On the Yang dataset, GAMED achieved a remarkable accuracy of 98.46%, surpassing the state-of-the-art MCNN by 2.16%. Meanwhile, GAMED also

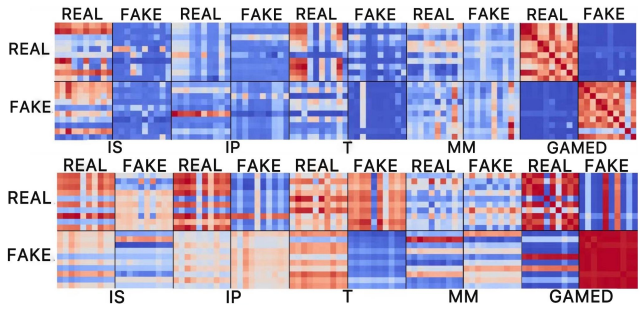

Figure 4: Heatmaps of cosine similarity on Fakeddit and Yang. Each heatmap cell shows the pairwise cosine similarity between the 64-dimensional representation from the coarse predictor of four modules and the whole model.

ranks first in Precision, Recall, and F1 on both Fakeddit and Yang. We tested recently developed BMR architecture on both Fakeddit and Yang datasets, and the results demonstrate that our GAMED outperforms BMR in almost all evaluation metrics, except for the Precision of Fakeddit.

Given the popularity of large language models (LLMs), in Table 2, we compare GAMED with LLMs-based fake news detection schemes, and the experiments are conducted on Fakeddit and depicted in Table 2. We obtained similar conclusions on the Yang dataset too. The Direct approach employs the model for fake news detection without any preprocessing of the input data, relying solely on the model's internal knowledge to directly generate predictions and reasoning. In contrast, the Chain of Thought (CoT) [63] approach enhances the model's ability to handle complex tasks by prompting it to "think step by step", guiding the model to first produce a reasoning process before delivering a final prediction. Both LLaVA [35] and the GPT-4 family [1] underperformed on this task, falling short of most traditional language models listed in Table 1. Furthermore, fine-tuned LLMs like InstructBLIP [11] or tool-enhanced LLMs like FacTool [9] did not improve the capabilities of LLMs. The state-of-the-art architecture LEMMA [65] has an accuracy of only 82.4%, i.e., significantly lower than our GAMED by 11.5%. In Precision, Recall and F1 score, GAMED shows improvement better than LLMs, except that it ranks second in Recall for the real news category, slightly lower than GPT-4V using CoT.

Our model performs better than the strong comparative models with several reasons. As mentioned before, we address some of the key shortcomings in the existing models. By employing modal decoupling and cross-modal synergy, GAMED preserves and enhances the discriminative features of each modality. The feature refinement components dynamically emphasize the most relevant information, in contrast to the static methods used before. Moreover, GAMED exploits semantic knowledge from pre-trained models, deepening its understanding of facts and relationships, which is a capability that many comparative models lack. Finally, our novel veto voting mechanism, which combines consensus and confidence, ensures that the most reliable predictions drive the final decision, offering greater flexibility than traditional voting or fusion methods.

As depicted in Figure 4, we randomly select ten fake and ten real news samples to visualize the heat map. The colours of the heat map

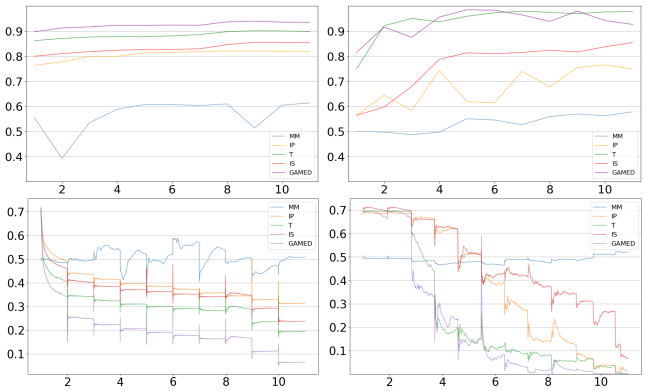

Figure 5: The comparison of accuracy (first row) and loss (second row) illustrates the learning curves of GAMED and its four modules during training. GAMED, IP, IS, and T exhibit ideal training trajectories, while MM shows minimal changes in both accuracy and loss. The training set on the left column and right column is Fakeddit and Yang, respectively.

range from dark blue (low similarity) to dark red (high similarity) showing the degree of similarity between features. The fake news features within a cell are more scattered than the real news features, which is often the reason for the difficulty in detection. However, GAMED successfully captures these discriminative news features. The differences between modules reflect that each module can provide distinctive perspectives and information when working. This again verifies our intuition that leveraging the discriminability and distinctiveness from modal decoupling can enhance the model's overall detection performance.

Figure 5 illustrates the learning curves of accuracy and loss for GAMED and its four modules over 11 epochs of training on the Fakeddit and Yang datasets. The learning curves on the Fakeddit dataset show steady improvement, with accuracy gradually increasing and loss sharply decreasing before stabilizing, indicating efficient learning and convergence. In contrast, the Yang dataset exhibits more variability, likely due to its smaller size and higher diversity and complexity of samples, which makes it more challenging for the model to consistently capture and learn patterns. The accuracy curves fluctuate before stabilizing, and although the loss initially decreases, it also shows greater volatility, reflecting the model's difficulty in consistently optimizing when confronted with diverse characteristics of the dataset. Despite these fluctuations, both datasets demonstrate that the models are learning and improving. Notably, although there are some fluctuations during training, the fusion module (MM) shows a generally stable trend in both accuracy and loss across these datasets, with no substantial improvements. This validates our view that the progress of the overall architecture is primarily driven by the synergy of multiple modalities, particularly the unimodal distinctiveness and discriminative power, while early fusion contributes only minimally.

4.3 Ablation Study

Removing Individual Modules: In Table 3, we present the ablation results. By removing modules, we find that the text module performs the best among all individual modules, achieving an accuracy of 90.1%, which even exceeds many excellent fake news detection models. Next are IS and IP, but their combined performance only reached 88.5%, still trailing the text module by 1.6%. The worst-performing individual module is MM, with a peak accuracy of only 61.4%, demonstrating the effectiveness of our weakened fusion module design. However, the combination of individual IS and T using the same data achieved an accuracy of 90.9%, significantly outperforming the MM module by 29.5%. More importantly, GAMED maintained strong detection capabilities despite the poor performance of the MM module. This further supports our view that the contributions of unimodal distinctiveness and their cross-modal synergies to model performance outweigh standalone modal fusion. Finally, compared with the high accuracy of GAMED, these removals prove that no single modality or any combination can reach the overall performance of GAMED.

Test	Accuracy
MM	0.614
IP	0.820
IS	0.855
IS+IP	0.885
T	0.901
IP+T	0.906
IS+T	0.909
IS+IP+T	0.914

Test	Accuracy
MM-only ComputeStats	0.884
w/o AdaIN	0.908
w/o Coarse Prediction	0.916
Replace IRNv2 with Inception-v3	0.924
Replace MMoE-Pro with ViT block	0.926
w/o Consistency Learning	0.926
Replace ERNIE2.0 with BERT	0.927
Replace MMoE-Pro with MMoE	0.928
Replace MAE-ViT with ViT	0.931
w/o Veto Voting	0.933
Replace ERNIE2.0 with ERNIE1.0	0.936
GAMED	**0.939**

Table 3: Ablation results of key modules and components in the GAMED design, tested on Fakeddit.

Knowledge Enhancement: We used BERT to replace ERNIE to process the text data, but the result dropped by 1.2%. Although both two models are based on similar transformer architecture, the advantage of ERNIE is that it further incorporates a structured knowledge graph to enhance the understanding of facts and relations. In addition, we used ERNIE1.0 instead of ERNIE2.0, and the result dropped by 0.3%. This is because ERNIE2.0 introduced a more complex knowledge increment strategy and larger knowledge parameters than ERNIE1.0 during pre-training. In the previous individual module removal, we found that the performance of a single text module was stronger than the combined performance of images modules, and the initial performance gap between IP and IS was large; but after the enhancement of the text module, not only the performance was greatly improved, both exceeded 90%; however, the gap between IP and IS became very small. ERNIE's victory highlights that external knowledge integration is crucial to enhancing the overall performance in complex tasks, i.e., multimodal fake news detection. In addition, to measure the impact of different feature extractors, we used InceptionV3 instead of Inception-ResNet-v2 and ViT instead of MAE-ViT, which resulted in a 1.5% and 0.8% drop in accuracy for GAMED, respectively.

Expert Network: We replaced MMoE-Pro with standard MMoE, resulting in a 1.1% drop in GAMED's accuracy. This decline is due to MMoE-Pro's enhancements in feature sharing and task relationship modelling, which allow the model to effectively process and fuse multimodal data. We replaced the MMoE-Pro networks with the ViT blocks for feature refinement, but the accuracy dropped

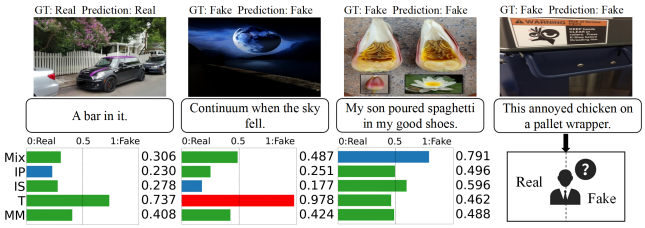

Figure 6: The interpretability of GAMED is illustrated through the decision-making process. The images and text are sourced from real examples in the Fakeddit test set. The first three examples from left to right represent decisions made by our GAMED, while the fourth example simulates decisions from other black-box models.

by 1.3%. This suggests that sharing information is beneficial for enhancing representations in our task, while the ViT blocks reduced the model's ability to flexibly and dynamically select the most relevant features for the task. We removed the coarse prediction step and the accuracy of GAMED dropped by 2.3%, which proves that the constraint of expert opinion is useful for evaluating the importance of modalities.

Adaptive Adjustment: We experimented with the mean and standard deviation of MM as the input of AdaIN to adjust the distribution of other unimodal modules. We found that the accuracy of GAMED dropped by 5.5%, which was a significant drop. In contrast, when we removed the consistency learning, the accuracy of GAMED dropped by 1.3%. This suggests that although consistency learning can improve the model's ability to identify fake news, prioritizing it does not bring greater benefits. Instead, it undermines the contribution of unimodal discriminability to the overall performance. We then removed the AdaIN setting for all modalities, and the accuracy of GAMED dropped by 3.1%. This is because AdaIN can adaptively adjust the feature distribution of different modalities to ensure that the most valuable features are provided for the interaction link.

Decision Making: We used concatenation-based late fusion instead of veto voting, resulting in a 0.6% drop in the accuracy of GAMED. This is because our carefully designed cross-modal interaction rules for veto voting can dynamically adjust the key modalities to ensure that the most reliable predictions have the greatest impact on the final decision. At the same time, other methods lose such flexibility and are more susceptible to noise.

4.4 Qualitative Analysis

As shown in Figure 6, we demonstrate the interpretability of the decisions made by GAMED using real examples. We randomly selected three prediction results on the Fakeddit test set and traced the corresponding samples. From left to right, following the veto voting rule in Section 3, the prediction result of the first sample is "Real", this is because the confidence of several modalities is between the pre-set low threshold θ_{low} and the high threshold θ_{high}, so the initial concatenated prediction P_{mix} is directly used as the final prediction. P_{mix} represents the result after the multimodal features are concatenated, and it makes full use of the complementarity of each modality to make the prediction result more comprehensive and reliable. The second sample shows that since the output confidence of the text module P_t is higher than the high threshold θ_{high} and higher than P_{mix}, the prediction of the initial P_{mix} is replaced with the prediction P_t as the final decision. The text modality provides extremely reliable information in this context. We can maximize the use of this reliable information and improve the accuracy of the final decision. The third sample reflects the decision made by GAMED when the output confidence of the image pattern module P_{ip} is lower than the low threshold θ_{low}. This is because too low confidence means that the information of this modality may be unreliable or misleading, even if it belongs to the majority class. By ignoring this unreliable prediction and comprehensively reconsidering the combination of the highest output in all modalities and the concatenated output, the robustness of the final decision is ensured. The fourth sample simulates the black-box decision-making process of many current models, in which the model cannot clearly explain the specific reasons for its decision. This may not only lead to a decrease in user trust in the model's prediction results but also make it difficult to debug and improve effectively when errors occur. In addition, GAMED's modal-decoupling design is also used for interpretability. For example, when AdaIN adaptively adjusts the feature distribution of different modules, we can judge the contribution of each modality and its discriminative features to the prediction. In contrast, those black-box models do not have a clear explanation path when processing input data. This means that the internal working mechanism of the model is invisible to users and developers, resulting in people being unable to understand or verify the reasoning process behind it even if the model makes a correct classification.

5 Conclusions

This paper developed GAMED – a novel architecture that significantly improves fake news detection. GAMED overcomes the shortcomings of current multimodal approaches through a dynamic mechanism of modal decoupling and cross-modal synergy. It embeds the benefits of semantic information encoded in knowledge graphs into the whole workflow from pre-trained language models. Feature selection is performed jointly by a mixture of experts, accompanied by subsequent adaptive distribution adjustment, progressively refining the feature representation of each modality in the pipeline and enhancing its discriminability and distinctiveness. Finally, a flexible and transparent decision process is introduced. Our experiments on benchmark datasets, Fakeddit and Yang, show that GAMED improves upon recent top-performing models in detection accuracy. Future work would explore adding more modalities such as audio or video for a more holistic analysis of fake news.

6 Acknowledgment

This work was supported by the Alan Turing Institute/DSO grant on improving multimodality misinformation detection through affective analysis. We gratefully acknowledge NVIDIA for providing computational resources through its NVIDIA Academic Hardware Grant Program 2021. Additional support was provided by the Interdisciplinary Research Pump-Priming Fund, University of Southampton.

7 Ethical Considerations

Some of the key ethical considerations include addressing the issues when models can perpetuate existing societal biases if the training data is biased. This can lead to discriminatory outcomes, such as unfairly targeting certain groups or individuals. Besides that, different cultures have varying norms and understandings of truth and fake news. Models trained on data from one culture may not perform well or ethically in another. Another fundamental challenge lies in overly aggressive detection models that could lead to the suppression of legitimate speech, particularly for marginalized voices or those critical of authority. Incorrectly flagging accurate information as fake news can damage reputations and stifle public discourse. Our goal in this work is to develop a model that could understand how it reached its conclusions. Black-box models make it difficult to identify and address biases. Overreliance on automated detection systems could erode trust in traditional media and journalism.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Ahmed Aleroud and George Karabatis. 2020. Bypassing Detection of URL-based Phishing Attacks Using Generative Adversarial Deep Neural Networks. *Proceedings of the Sixth International Workshop on Security and Privacy Analytics* (2020).

[3] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2023. Towards COVID-19 fake news detection using transformer-based models. *Knowledge-Based Systems* 274 (2023), 110642.

[4] Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. 2021. Exmulf: an explainable multimodal content-based fake news detection system. In *International Symposium on Foundations and Practice of Security*. Springer, 177–187.

[5] K Anirudh, Meghana Srikanth, and A Shahina. 2023. Multilingual Fake News Detection in Low-Resource Languages: A Comparative Study Using BERT and GPT-3.5. In *International Conference on Speech and Language Technologies for Low-resource Languages*. Springer, 387–397.

[6] C. V. Asha, Akash Yadav, Dr. A. P Nirmala, Abhijit Sen, A Arvind Raj, and Abhinav Rajan. 2024. An Effective Assessment of Machine Learning Approaches for Fake News Detection. *2024 International Conference on Inventive Computation Technologies (ICICT)* (2024), 169–174.

[7] Belhassen Bayar and Matthew C Stamm. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2691–2706.

[8] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*. 2897–2905.

[9] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).

[10] Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining* 13, 1 (2023), 101.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv* abs/2305.06500 (2023).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

[13] Hui Fang, Chongcheng Chen, Yunfei Long, Ge Xu, and Yongqiang Xiao. 2022. DTCRSKG: A deep travel conversational recommender system incorporating knowledge graph. *Mathematics* 10, 9 (2022), 1402.

[14] Lifang Fu, Huanxin Peng, and Shuai Liu. 2023. KG-MFEND: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing* 79, 16 (2023), 18417–18444.

[15] Siva Charan Reddy Gangireddy, Deepak P, Cheng Long, and Tanmoy Chakraborty. 2020. Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM conference on hypertext and social media*. 75–83.

[16] Xingyu Gao, Xi Wang, Zhenyu Chen, Wei Zhou, and Steven CH Hoi. 2024. Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Transactions on Multimedia* (2024).

[17] Ankit Gautam, S Vijayakumar Bharathi, Dhanya Pramod, and Kanchan Patil. 2023. Fake textual and image news detection on social media using natural language processing. In *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*. IEEE, 1–6.

[18] Hao Guo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2023. Interpretable Fake News Detection with Graph Evidence. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 659–668.

[19] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. 2024. Siamese masked autoencoders. *Advances in Neural Information Processing Systems* 36 (2024).

[20] Mostafa Haghir Chehreghani. 2024. A Review on the Impact of Data Representation on Model Explainability. *Comput. Surveys* (2024).

[21] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Gener. Comput. Syst.* 117 (2021), 47–58.

[22] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113.

[23] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open* 3 (2022), 133–155.

[24] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 754–763.

[25] Jiaheng Hua, Xiaodong Cui, Xianghua Li, Keke Tang, and Peican Zhu. 2023. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing* 136 (2023), 110125.

[26] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1501–1510.

[27] Saurabh Jaiswal and Mr. Vivek Rai. 2024. Enhancing Generalisation In Fake News Detection: A Comparative Evaluation of Naïve Bayes and Random Forest Approaches. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)* (2024).

[28] Mingyue Jiang, Chang Jing, Liming Chen, Yang Wang, and Shouqiang Liu. 2024. An application study on multimodal fake news detection based on Albert-ResNet50 Model. *Multimedia Tools and Applications* 83, 3 (2024), 8689–8706.

[29] Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2023. Multimodal fake news detection via progressive fusion networks. *Information processing & management* 60, 1 (2023), 103120.

[30] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[31] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).

[32] Rina Kumari, Nischal Ashok, Pawan Kumar Agrawal, Tirthankar Ghosal, and Asif Ekbal. 2023. Identifying multimodal misinformation leveraging novelty detection and emotion recognition. *Journal of Intelligent Information Systems* 61, 3 (2023), 673–694.

[33] Batool Lakzaei, Mostafa Haghir Chehreghani, and Alireza Bagheri. 2024. Disinformation detection using graph neural networks: a survey. *Artificial Intelligence Review* 57, 3 (2024), 52.

[34] Jaeyoung Lee, Ximing Lu, Jack Hessel, Faeze Brahman, Youngjae Yu, Yonatan Bisk, Yejin Choi, and Saadia Gabriel. 2024. How to Train Your Fact Verifier: Knowledge Transfer with Multimodal Open Models. *ArXiv* abs/2407.00369 (2024).

[35] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2024).

[36] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. 2022. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems* 35 (2022), 14290–14302.

[37] Jia Li and Minglong Lei. 2022. A Brief Survey for Fake News Detection via Deep Learning Models. In *International Conference on Information Technology and Quantitative Management*.

[38] Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, Vol. 29. Association for Computational Linguistics, 7124–7135.

[39] Qi Liu, Yuanyuan Jin, Xuefei Cao, Xiaodong Liu, Xiaokang Zhou, Yonghong Zhang, Xiaolong Xu, and Lianyong Qi. 2024. An Entity Ontology-Based Knowledge Graph Embedding Approach to News Credibility Assessment. *IEEE Transactions on Computational Social Systems* (2024).

[40] Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion* 104 (2024), 102203.

[41] Qiang Lu, Xia Sun, Zhizezhang Gao, Yunfei Long, Jun Feng, and Hao Zhang. 2024. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. *Information Processing & Management* 61, 1 (2024), 103538.

[42] Muhammad Luqman, Muhammad Faheem, Waheed Yousuf Ramay, Malik Khizar Saeed, and Majid Bashir Ahmad. 2024. Utilizing ensemble learning for detecting multi-modal fake news. *IEEE Access* (2024).

[43] Alex Munyole Luvembe, Weimin Li, Shaohau Li, Fangfang Liu, and Xing Wu. 2024. CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management* 61, 3 (2024), 103653.

[44] Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2023. Kapalm: Knowledge graph enhanced language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3999–4009.

[45] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1930–1939.

[46] Priyanka Meel and Dinesh Kumar Vishwakarma. 2021. HAN, image captioning, and forensics ensemble multimodal fake news detection. *Inf. Sci.* 567 (2021), 23–41.

[47] Amit Praseed, Jelwin Rodrigues, and P Santhi Thilagam. 2023. Hindi fake news detection using transformer ensembles. *Engineering Applications of Artificial Intelligence* 119 (2023), 105731.

[48] 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)*. IEEE, 518–527.

[49] Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and BRNN. In *Proceedings of the 3rd international workshop on rumours and deception in social media (RDSM)*. 14–31.

[50] Anne Schulz, Richard Fletcher, and Rasmus Kleis Nielsen. 2024. The role of news media knowledge for how people use social media for news in five countries. *new media & society* 26, 7 (2024), 4056–4077.

[51] Jae-Seung Shim, Yunju Lee, and Hyunchul Ahn. 2021. A link2vec-based fake news detection model using web search results. *Expert Systems with Applications* 184 (2021), 115491.

[52] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 395–405.

[53] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. 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. 557–565.

[54] Chenguang Song, Yiyang Teng, Yangfu Zhu, Siqi Wei, and Bin Wu. 2022. Dynamic graph neural network for fake news detection. *Neurocomputing* 505 (2022), 362–374.

[55] Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and S Yu Philip. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (2023), 12736–12749.

[56] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8968–8975.

[57] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[58] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[59] Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval* 12, 2 (2023), 28.

[60] Pawan Kumar Verma, Prateek Agrawal, Vishu Madaan, and Radu Prodan. 2023. MCred: multi-modal message credibility for fake news detection using BERT and CNN. *Journal of Ambient Intelligence and Humanized Computing* 14, 8 (2023), 10617–10629.

[61] Bin Wang, Yong Feng, Xian-cai Xiong, Yong-heng Wang, and Bao-hua Qiang. 2023. Multi-modal transformer using two-level visual features for fake news detection. *Applied Intelligence* 53, 9 (2023), 10429–10443.

[62] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[64] 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.

[65] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. LEMMA: Towards LVLM-Enhanced Multimodal Misinformation Detection with External Knowledge Augmentation. *arXiv preprint arXiv:2402.11943* (2024).

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

[67] Hongyu Yang, Jinjiao Zhang, Liang Zhang, Xiang Cheng, and Ze Hu. 2024. MRAN: Multimodal relationship-aware attention network for fake news detection. *Computer Standards & Interfaces* 89 (2024), 103822.

[68] 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).

[69] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 37. 5384–5392.

[70] Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. 2025. Learning the Unlearned: Mitigating Feature Suppression in Contrastive Learning. In *European Conference on Computer Vision*. Springer, 35–52.

[71] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.

[72] 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*. Springer, 354–367.

[73] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2825–2830.

[74] Melissa Zimdars and Kembrew McLeod. 2020. *Fake news: Understanding media and misinformation in the digital age*. MIT Press.