Towards Content Authenticity: Multimodal Fake News Detection and AI-Generated Text Identification

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science

by

Nidhi Gupta
Delhi Technological University
Bachelor of Technology in Software Engineering, 2020

August 2025
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

Qinghua Li, Ph.D.
Thesis Advisor and Committee Chair

Lu Zhang, Ph.D.
Committee Member

Brajendra Nath Panda, Ph.D.
Committee Member

ABSTRACT

In today's digital world, the spread of fake news and the rise of AI-generated text have become major threats to content authenticity and public trust. This thesis addresses both challenges through two complementary research directions: detecting fake news using multimodal features, and identifying AI-generated text using semantic and structural reasoning. The first part of the work focuses on fake news detection by introducing a novel model that combines text and image features through a unique rotational attention mechanism. Unlike traditional attention methods, this approach rotates the roles of query, key, and value across modalities to capture deeper interactions. Additionally, the model incorporates external domain information by linking news posts to top-ranked websites from Google search results, which helps assess the credibility of content based on its broader web context. This results in a more reliable and accurate fake news detection system that outperforms existing state-of-the-art methods. The second part presents SGG-ATD, a new framework for detecting AI-generated text. It uses masked language modeling to measure sentence coherence, followed by constructing a graph where keywords—both original and predicted—are connected based on semantic and contextual similarity. A Graph Convolutional Network (GCN) is then used to learn structural relationships within the text for final classification. Experimental results demonstrate that SGG-ATD achieves high F1-scores and consistently outperforms strong baselines. This method contributes to robust AI text detection, supporting accountability and resilience against AI-driven misinformation.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Qinghua Li, for their continuous guidance, encouragement, and support throughout this research journey. Their insights, feedback, and mentorship have been invaluable to both the technical depth and clarity of this thesis.

I would also like to thank the faculty and staff of the Department of Electrical Engineering & Computer Science at the University of Arkansas for providing a supportive academic environment and the resources necessary to carry out this work. Special thanks to my committee members, Dr. Lu Zhang and Dr. Brajendra Nath Panda, for their time and thoughtful suggestions that helped refine this research.

I am grateful to my peers and labmates for their stimulating discussions, collaboration, and camaraderie, which made the research process both intellectually enriching and enjoyable.

A heartfelt thank you to my family—especially my husband and parents—for their unwavering love, patience, and belief in me. Their support has been my foundation throughout this academic journey. I am also grateful to my friends and siblings who stood by me during moments of stress and shared in moments of joy.

Lastly, I would like to acknowledge the inspiration that comes from working on research that aligns with the greater purpose of promoting truth, accountability, and trust in digital information systems. It is my hope that this work contributes meaningfully to that vision.

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS

Nidhi Gupta, Qinghua Li, and Lu Zhang "CAMFeND: Credibility-Aware Multimodal Fake News Detection with Rotational Attention", Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2025 (accepted). Referenced in Chapter 2.

Nidhi Gupta and Qinghua Li, "Seeing Through the Mask: AI-Generated Text Detection with Similarity-Guided Graph Reasoning", (in review). Referenced in Chapter 3.

1 Introduction

In the digital age, the integrity of information is under unprecedented threat. With the widespread adoption of social media platforms and real-time content sharing [1], the speed at which information travels has vastly outpaced traditional verification mechanisms. This ecosystem has become fertile ground for the spread of false, misleading, or synthetically generated information that can influence public opinion [2], manipulate social movements, and erode trust in institutions. As technology continues to evolve, so do the techniques for generating and distributing deceptive content—making it harder to distinguish truth from fabrication in the digital space.

One of the most alarming manifestations of this problem is fake news, which leverages persuasive writing, emotional imagery, and sometimes partial truths to deceive readers [3]. Fake news not only misinforms individuals but also has far-reaching consequences on democratic processes, public health decisions, and societal harmony. Compounding this issue is the rise of AI-generated text, which is now capable of producing content that closely mimics human language, tone, and style. Models like GPT-4 and similar large language models (LLMs) [4] have made it possible to create high-quality synthetic articles, reviews, or comments that appear authentic to both humans and traditional content filters.

These two phenomena—fake news and AI-generated content—represent different sides of the same problem: the loss of content authenticity and credibility in an increasingly automated and interconnected world. While fake news focuses on the deliberate spread of misinformation, AI-generated text introduces the risk of unintentionally or maliciously produced synthetic content that may not be explicitly false but is still artificially authored and potentially manipulative. Together, they pose a complex and evolving challenge that traditional machine learning approaches, often based on surface-level features, struggle to address.

To address the growing threat to information integrity, it is essential to develop intelligent systems capable of assessing both the credibility and origin of digital content. This requires moving beyond simple keyword-based or rule-driven methods and adopting more nuanced, context-aware, and model-informed strategies. In this thesis, we focus on two critical and complementary challenges in the broader effort to detect deceptive content: multimodal fake news detection and AI-generated text detection. Both tasks contribute to the overarching goal of identifying and mitigating synthetic or misleading information in the digital space. For fake news detection, we design methods that jointly leverage textual, visual, and domain-level credibility cues to assess the veracity of news posts. For AI-generated text detection, we address the increasingly difficult task of distinguishing human-written content from that produced by large language models, with a focus on capturing semantic coherence. To this end, we propose deep learning frameworks that integrate attention-based multimodal fusion, graph-based reasoning, and similarity-guided inference. This thesis is structured around these two components, each presenting a novel framework tailored to its task while contributing to the goal of strengthening content authenticity in digital environments.

1.1 Fake News Detection

Early fake news detection models primarily focused on analyzing textual content using linguistic features, syntax patterns, or stylistic cues [5]. While effective to some extent, these single-modal approaches often failed to capture the full context of misinformation. As fake news began to rely more heavily on emotionally provocative images to enhance believability, research shifted toward multimodal detection frameworks that leverage both text and visual information. Despite this advancement, two key limitations remain.

First, many approaches exhibit limited cross-modal interaction, using static alignment techniques such as co-attention [6] or relationship-aware attention [7], which fail to capture the dynamic and evolving dependencies between text and image features. Second, most models neglect the credibility of the news source, treating content in isolation without considering domain-level trustworthiness.

To address these gaps, Chapter 2 of this thesis proposes a novel architecture that introduces a rotational attention mechanism, which dynamically rotates the roles of query, key, and value across modalities—enabling richer, bidirectional interaction between text and image features [8]. Additionally, the model incorporates news domain credibility by associating news posts with top-ranked domains retrieved through web search, thus grounding the content in contextual reliability. This combined framework enhances both the depth of multimodal fusion and the robustness of credibility reasoning, achieving superior performance on benchmark fake news datasets.

1.2 AI-Generated Text Detection

Existing AI detection approaches typically rely on surface-level indicators such as token likelihoods [9], and statistical irregularities [10]. While these methods offer reasonable performance in controlled settings, they struggle to generalize across different writing styles, domains, and prompt variations. Moreover, most techniques treat each text sample in isolation, overlooking the deeper structural and semantic patterns that characterize human versus machine-generated language.

Chapter 3 of this thesis presents a new approach that addresses these challenges by incorporating contextual and structural reasoning into the detection process. Rather than focusing solely on local features, the proposed method captures broader semantic coherence and relationships within the text, allowing it to better distinguish the subtle regularities and predictability often found in AI-generated content. This enables more robust performance across a variety of content types, including essays, news articles, technical descriptions, and creative narratives—ultimately advancing the goal of trustworthy AI and content verification.

1.3 Toward a Broader Effort for Content Authenticity

Together, the two core components of this thesis contribute to a broader perspective on digital content authenticity by addressing the recent challenges of fake news detection and AI-generated text identification. While the first focuses on multimodal features and external knowledge (via news domains), the second focuses on textual coherence and similarity patterns using graph-based reasoning. The shared goal is to move toward AI systems that can contextually understand, verify, and interpret content in an environment where deception is scalable and increasingly machine-powered.

This thesis, therefore, not only contributes novel architectures in each component, but also lays the groundwork for future research on integrating these detection strategies into real-world content verification pipelines—helping societies better navigate the evolving landscape of truth and fabrication in the digital age.

Bibliography

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] N. Gupta, Q. Li, and L. Zhang, "Camfend: Credibility-aware multimodal fake news detection with rotational attention," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[9] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps *et al.*, "Release strategies and the social impacts of language models," *arXiv preprint arXiv:1908.09203*, 2019.

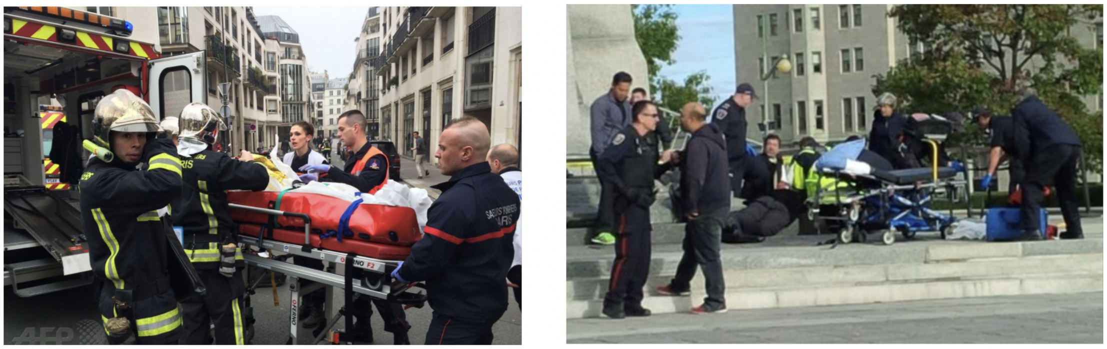
[10] G. Jawahar, M. Abdul-Mageed, and L. V. Lakshmanan, "Automatic detection of machine generated text: A critical survey," *arXiv preprint arXiv:2011.01314*, 2020.

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

2.1 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 2.1), 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.

Early detection approaches [2, 3, 4, 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:



10 dead as shots fired at French satirical weekly

Canadian soldier shot in Ottawa a reservist from Hamilton

Figure 2.1: Illustrations of fake news stories sourced from the Pheme [1] dataset.

- **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.

2.2 Related Work

2.2.1 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, 21, 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.

2.2.2 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.

2.2.3 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, 44, 45]. Relationship-aware attention, co-attention, and knowledge-augmented features were further advanced by work like [15], [46, 47, 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.

2.3 Method

In this section, we present our proposed multimodal fake new detection framework as illustrated in Figure 2.2, 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.

2.3.1 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.

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 $1 \times S$.

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, $\mathbf{h}_i^{(0)} \in R^{d_{text}}$, while the news domain nodes (v_j) are initialized with Word2Vec [49] embeddings, $\mathbf{h}_j^{(0)} \in R^{d_{domain}}$.

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:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}_A \mathbf{h}_i^{(l)} \parallel \mathbf{W}_B \mathbf{h}_j^{(l)}\right]\right) \tag{2.1}$$

where $\mathbf{W}_A \in R^{d \times d_{text}}$ and $\mathbf{W}_B \in R^{d \times d_{domain}}$ are learnable weight matrices specific to the two node types, $\mathbf{a} \in R^{2d}$ 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:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_A(i)} \exp(e_{ik})} \tag{2.2}$$

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

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}:

$$\mathbf{h}_i^{(l+1)} = \sigma \left(\sum_{j \in \mathcal{N}_A(i)} \alpha_{ij} \mathbf{W}_B \mathbf{h}_j^{(l)} \right) \tag{2.3}$$

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 $v_i \in \mathcal{V}_A$:

$$\mathbf{h}_j^{(l+1)} = \sigma \left(\sum_{i \in \mathcal{N}_B(j)} \alpha_{ji} \mathbf{W}_A \mathbf{h}_i^{(l)} \right) \tag{2.4}$$

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.

2.3.2 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.

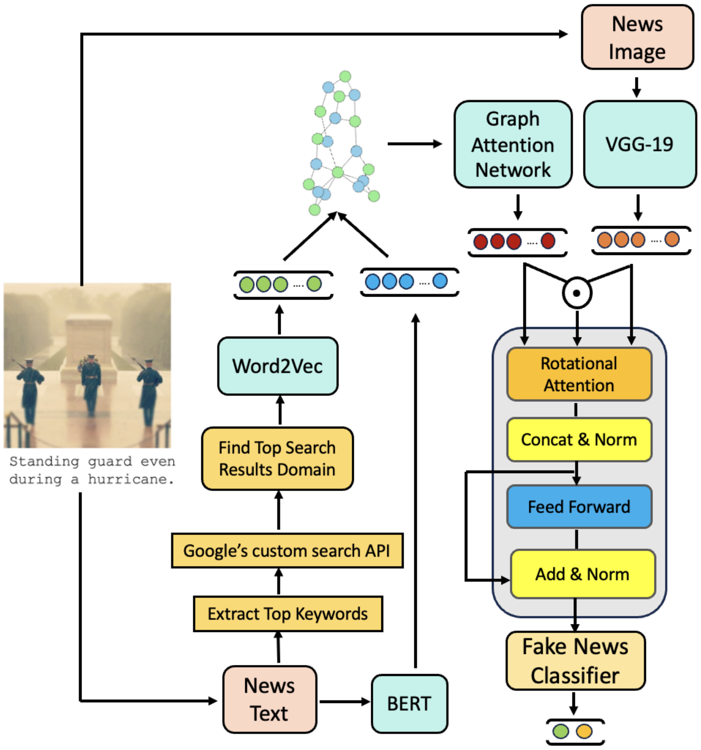


Figure 2.2: Architecture of our CAMFeND model. Text features from BERT are enhanced using a Graph Attention Network capturing news post-domain relationships, while visual features come from VGG-19. A rotational attention mechanism exchanges query, key, and value roles between GAT and VGG-19 embeddings. 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.

2.3.3 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.

Traditional Attention Mechanism

The standard multi-head self-attention (MSA) [52] block shown in Figure 2.3(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:

$$\mathrm{A}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{2.5}$$

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

The Multi-Head Attention is calculated as:

$$h_j = \mathrm{A}(QW_j^Q, KW_j^K, VW_j^V) \tag{2.6}$$

$$\mathrm{MHA}(Q, K, V) = \mathrm{concat}(h_1, \ldots, h_m)W^O \tag{2.7}$$

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

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

$$\mathrm{FFN}(x) = \max(0, xW_1)W_2 \tag{2.8}$$

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

Rotational Attention Mechanism

The rotational attention mechanism in Figure 2.3(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.

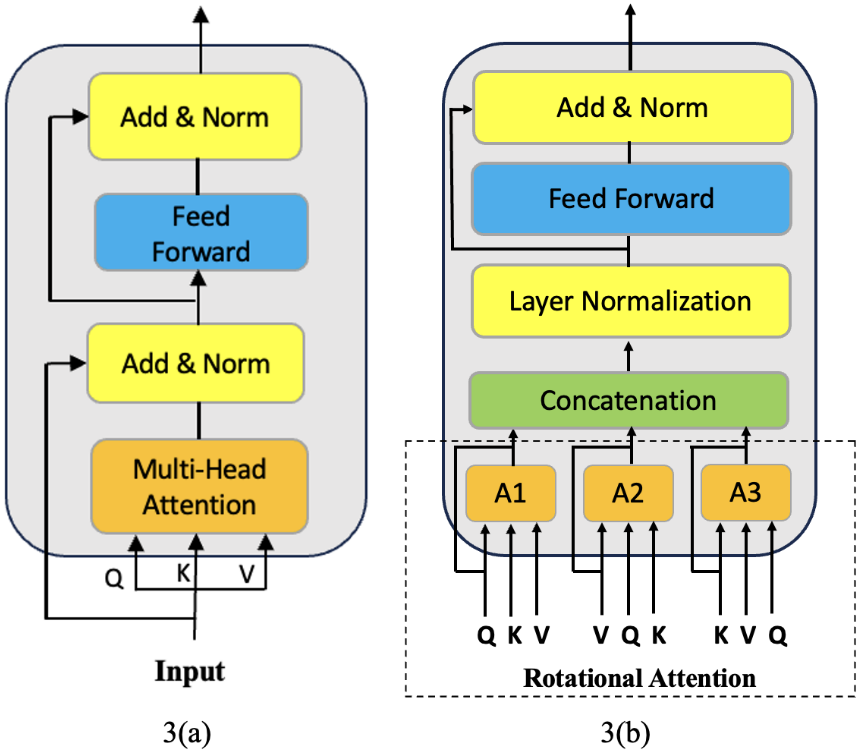


Figure 2.3: (a) Self Attention and (b) Rotational Attention: Q, K, and V roles rotate across three attention layers.

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 approach. The rotational attention mechanism proceeds as follows:

Attention 1

$$\mathbf{A}_1 = \mathrm{A}(\mathbf{I}_{vgg}, \mathbf{T}_{gat}, \mathbf{I}_{vgg} \odot \mathbf{T}_{gat}) + \mathbf{I}_{vgg} \tag{2.9}$$

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, $\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}$.

Attention 2

$$\mathbf{A}_2 = \mathrm{A}(\mathbf{I}_{vgg} \odot \mathbf{T}_{gat}, \mathbf{I}_{vgg}, \mathbf{T}_{gat}) + \mathbf{I}_{vgg} \odot \mathbf{T}_{gat} \tag{2.10}$$

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 $\mathbf{T}_{gat}$.

Attention 3

$$\mathbf{A}_3 = \mathrm{A}(\mathbf{T}_{gat}, \mathbf{I}_{vgg} \odot \mathbf{T}_{gat}, \mathbf{I}_{vgg}) + \mathbf{T}_{gat} \tag{2.11}$$

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 $\mathbf{I}_{vgg}$.

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:

$$\mathbf{A}_{concat} = [\mathbf{A}_1; \mathbf{A}_2; \mathbf{A}_3] \tag{2.12}$$

$$\mathbf{A}_{norm} = \mathrm{LayerNorm}(\mathbf{A}_{concat}) \tag{2.13}$$

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.

$$\mathbf{A}_{ff} = \mathrm{FFN}(\mathbf{A}_{norm}) \tag{2.14}$$

$$\mathbf{A}_{final} = \mathrm{LayerNorm}(\mathbf{A}_{ff} + \mathbf{A}_{norm}) \tag{2.15}$$

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.

2.3.4 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:

$$\hat{y}_k = \sigma(\max(0, W_c \cdot \mathbf{A}_{finalK})W_s) \tag{2.16}$$

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:

$$\mathcal{L}(\Theta) = -\sum_{k=1}^{N}[Y_k \log(\hat{y}_k) + (1 - Y_k)\log(1 - \hat{y}_k)] \tag{2.17}$$

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

2.4 Evaluations

2.4.1 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 2.1 shows the dataset statistics.

Table 2.1: Data statistics for two real-world datasets.

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

2.4.2 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 $d_{\mathrm{ff}} = 512$. 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.

2.4.3 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.

Table 2.2: 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**

- **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.

Table 2.3: 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**

- **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 2.2 and 2.3 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.

2.4.4 Ablation Results and Discussions

Table 2.4 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 2.4: 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**

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.

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.

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.

2.5 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.

Bibliography

[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.

3 Seeing Through the Mask: AI-Generated Text Detection with Similarity-Guided Graph Reasoning

3.1 Introduction

In an era where machines write as fluently as humans, we are entering a new chapter in how information is produced, consumed, and trusted. Large Language Models (LLMs) such as GPT-4 [1], Claude [2], and LLaMA [3] have made it nearly effortless to generate essays, news articles, reviews, and even research papers with human-like fluency. What was once an imaginative leap—a machine composing coherent and contextually accurate paragraphs—is now commonplace. The boundary between synthetic and authentic language is becoming indistinguishable to the naked eye.

As this generative capability becomes more accessible and widespread—through models like GPT [4], BERT [5], and T5 [6]—its applications have expanded rapidly to include content creation, conversational agents, and real-time translation [7, 8]. However, this growing realism brings profound challenges: from misinformation and fake news propagation to academic dishonesty and erosion of digital trust [9, 10, 11]. With AI-generated content becoming nearly indistinguishable from human writing, questions around authorship, authenticity, and accountability are now more urgent than ever.

As these models seamlessly blend into communication workflows, a new and urgent challenge emerges. Educators, journalists, policymakers, and even AI developers are increasingly grappling with a pressing question: How do we determine who—or what—authored a piece of text? From student assignments generated at the push of a button to fabricated news articles and automated spam campaigns, the misuse of LLMs has already begun to erode trust in written communication. Worse still, existing detection techniques are rapidly losing ground. Conventional methods such as [12, 13] typically rely on shallow linguistic heuristics, statistical features, or supervised classifiers trained on outputs from known language models. While these approaches show promise on curated benchmarks, they often struggle to generalize across domains or withstand adversarial rewriting, paraphrasing, and stylistic obfuscation [14, 15]. As a result, adversaries can easily manipulate AI-generated text to appear convincingly human. This underscores the need for detection frameworks that move beyond surface-level patterns and engage with the structural underpinnings of language.

However, most existing detection methods fail to operationalize this structural perspective. Despite recent advances, two major limitations persist:

- *Lack of structural reasoning:* While prior work recognizes that AI-generated text tends to exhibit higher predictability, many existing methods rely only on surface-level cues such as per-token probabilities [16, 17] or shallow statistical features [12, 18], failing to model the deeper contextual and compositional structures that give rise to these patterns.

- *Limited generalization across varied domains:* Existing detectors such as DetectGPT [14] and Ghostbuster [15] often underperform when applied to unseen domains or writing styles.

At the heart of this dilemma lies a deeper question—not just whether a piece of text is AI-generated, but whether its structure and predictability reveal traces of its origin. Human language, while flexible and expressive, carries with it natural irregularities and subtleties rooted in reasoning, creativity, and intent. AI-generated text, by contrast, is often more formulaic, exhibiting higher token-level predictability and stylistic consistency. Capturing this difference requires methods that can perceive and represent the interplay between meaning, context, and linguistic structure.

Building on this intuition, we propose a new approach to AI-generated text detection that leverages masked language modeling to uncover patterns of semantic coherence and contextual regularity. We first extract content-rich keywords from the input text and mask a subset of them. A pretrained language model predicts the masked keywords, and both the extracted and predicted keywords are used to construct a contextual graph. In this graph, nodes represent keywords, and edges encode lexical semantics and contextual similarity. This structure allows our framework to reason over meaning-based patterns and generative signals—enabling more accurate and robust classification.

Our method, AI-Generated Text Detection with Similarity-Guided Graph Reasoning (SGG-ATD)—addresses the limitations outlined earlier by combining masked language modeling with graph-based reasoning:

- We construct a graph that connects original keywords and LLM-predicted keywords, allowing the model to capture how words relate in both meaning and context. This enables the model to move beyond isolated word-level analysis and instead reason over the structural and contextual flow of the text—an area where AI-generated writing often differs from human-authored content.

- We enhance the model's ability to generalize across varied text types—such as news articles, essays, technical descriptions, and creative writing—by using masked keyword prediction. This approach helps the model learn the underlying predictability and structure of a passage, enabling it to identify generative patterns that persist across different domains and writing styles.

By combining semantic meaning and LLM-prediction patterns in a graph structure, SGG-ATD provides a unified way to understand how words relate and how likely they are to appear in context. Unlike traditional models, our approach captures the deeper structure of how words connect and flow. This helps the model better recognize patterns that are typical of AI-generated content—even when the text is rewritten or comes from a different domain.

Empirical evaluations across multiple datasets — including news, creative writing, essays, and vulnerability descriptions — show that our framework outperforms strong baselines, achieving superior F1 scores and generalization across both in-distribution and out-of-distribution settings.

The remainder of this chapter is structured as follows: Section 3.2 presents related work, Section 3.3 outlines the proposed method, Section 3.4 provides a detailed evaluation and analysis of results, and Section 3.5 concludes this chapter with final insights.

3.2 Related Work

Large language models (LLMs) dramatically advanced the quality of machine-generated text, narrowing the gap with human writing across diverse domains. Early models like GPT-2 and GPT-3 demonstrated few-shot and zero-shot capabilities that pushed the frontier of language generation [19, 4]. These were later scaled further in models such as [20, 21], which showed that architectural and computational scale alone can yield significant performance improvements across instruction following, translation, and question answering tasks. Despite these capabilities, researchers also highlighted linguistic differences between LLM-generated and human text, such as reduced factuality or coherence in early generations in [18, 22].

To detect such content, many approaches were developed that analyzed surface-level features, probability metrics, or neural representations. For instance, [12] visualized token-level likelihoods to help humans distinguish AI-generated text, while [14] used log-probability curvature from perturbed inputs to separate model-written content from human-authored responses. Extending these ideas, [15] proposed a structured approach by scoring token probability distributions from weaker models. Models such as [23] combined DeBERTa and traditional classifiers, showing strong results in English web text. A common thread in these models was that their effectiveness often relied on access to scoring APIs or logit distributions, which may not be available for closed-source LLMs.

To move beyond token-level metrics, recent efforts incorporated structure and semantics. For example, [24] proposed a novel rewriting-based detection strategy, where text was passed through a rewriting model and the degree of transformation was used as a signal of authenticity. Similarly, [25] used graph neural networks to model word co-occurrence in texts and extracted deeper contextual patterns for detection. These approaches attempted to address the brittleness of detectors that relied only on shallow cues.

Domain generalization emerged as a critical challenge for detection models, especially when trying to flag content from unseen generators like GPT-4 or Claude. To tackle this, [26] proposed a framework that combined domain-adversarial learning and contrastive loss to generalize across LLMs without requiring retraining. Similarly, [27] formulated detection as a domain adaptation problem, allowing models trained on legacy LLMs to adapt to modern ones without labeled data. These approaches attempted to future-proof detectors against rapid advances in generation technologies.

In parallel, watermarking-based detection saw a resurgence. One line of work such as [28] introduced a soft watermark that biased generation toward a known token distribution, while [29] proposed a statistically robust watermark with provable guarantees under paraphrasing. A comprehensive survey [30] examined earlier watermarking efforts and highlighted challenges like multilinguality and visibility under adversarial attacks. These techniques offered post-hoc verifiability but depended on model-side cooperation.

Despite these developments, a growing body of work showed that many detectors were vulnerable to simple evasion techniques. Rephrasing, synonym replacement, and style-shifting can significantly reduce detection accuracy, even for strong models like [14] or [23]. Some attacks even worked across detectors by perturbing only the prompt without changing semantics, as shown in recent jailbreak studies [31, 32]. These findings raised concerns about the long-term robustness of detection systems.

Prompt engineering has also played a dual role—both in instructing models for tasks and in enabling or defeating detection. Chain-of-thought prompting, prefix tuning, and zero-shot reasoning enhanced reasoning fluency in LLMs [33, 34, 35]. However, these same mechanisms can be exploited to disguise AI-generated text or control its stylistic fingerprint as in [36].

Finally, questions of fairness and bias in detection remain largely underexplored. As [37] showed that existing detectors disproportionately misclassified non-native English writing as AI-generated, it raised concerns about fairness in academic or professional contexts. Simultaneously, societal studies like [38] showed that AI-generated content—while often helpful—differed in tone and formal structure, affecting its acceptability depending on the task.

Together, this body of work underscores that despite significant progress, AI-generated text detection remains challenging—particularly under adversarial, cross-domain, and stylistically diverse scenarios. In response, our framework shifts focus to the underlying structure and contextual predictability of the text by modeling relationships between original and LLM-predicted keywords. This alternative perspective aims to offer robustness in detection without relying on model-specific signatures.

While detection research progressed rapidly, many existing methods suffered from significant constraints. A large subset of detectors—including those based on log-probabilities or token distributions such as [14, 15]—relied on white-box access to the generating model, which was impractical for closed APIs or unseen LLMs. Others like [39, 23] depended on stylistic patterns or frequency-based features that can be evaded through prompt rephrasing or synonym substitution. Watermarking methods like [29, 28], while provably robust in controlled settings, require model cooperation can be vulnerable to transformations in real-world use. Even domain adaptation frameworks like [26, 40], though effective, still relied on feature alignment rather than deeper semantic grounding.

In contrast, our approach focuses on modeling the semantic and structural coherence of the text itself, independent of the generator's internal distribution. By leveraging graph-based representations constructed from original and LLM-predicted keywords, our method captures localized semantic relationships and models contextual predictability. This design encourages robustness against common challenges such as paraphrasing and domain variation, offering a detection strategy that does not rely on prior knowledge of the generator or labeled outputs.

3.3 Method

In this section, we present our proposed AI Text Detection Framework, SGG-ATD (Figure 3.1), which identifies AI-generated text by combining masked language modeling with graph-based reasoning. This novel framework captures semantic associations and contextual predictability through a context-enriched graph formulation. It comprises the following four key components:

1. **Keyword Extraction and Masking:** This module extracts syntactically meaningful keywords (nouns and verbs) using Part-of-Speech (POS) tagging [41]. To simulate partial context and expose latent structural cues, 30% of these keywords are randomly selected and replaced with the `<mask>` token.

2. **Masked Keyword Prediction:** The masked input text is then passed through a pre-trained ALBERT-base-v2 model [42], which predicts the missing keywords based on surrounding context. These predictions provide insight into keyword-level predictability, revealing structural regularities often present in AI-generated content.

3. **Graph Construction with Dual Similarity Encoding:** A graph is constructed where nodes represent original and LLM-predicted keywords. Edges are weighted using cosine similarity and contextual similarity, which are combined into a unified adjacency matrix for graph-based reasoning.

4. **Graph-Based Classification via Graph Convolutional Networks (GCN):** The

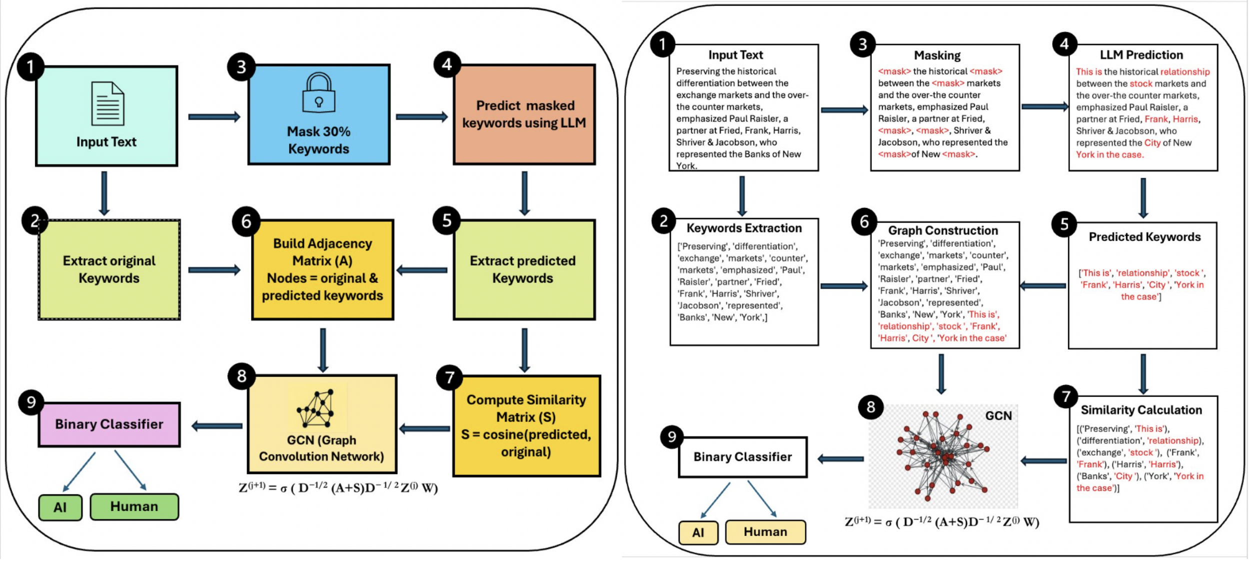

1
Input Text
3
Mask 30% Keywords
4
Predict masked keywords using LLM
2
Extract original Keywords
6
Build Adjacency Matrix (A) Nodes = original & predicted keywords
5
Extract predicted Keywords
9
Binary Classifier
8
GCN (Graph Convolution Network)
7
Compute Similarity Matrix (S) S = cosine(predicted, original)
AI
Human
$Z^{(j+1)} = \sigma (D^{-1/2} (A+S)D^{-1/2}Z^{(j)} W)$
1
Input Text
Preserving the historical differentiation between the exchange markets and the over-the counter markets, emphasized Paul Raisler, a partner at Fried, Frank, Harris, Shriver & Jacobson, who represented the Banks of New York.
3
Masking
<mask> the historical <mask> between the <mask> markets and the over-the counter markets, emphasized Paul Raisler, a partner at Fried, <mask>, <mask>, Shriver & Jacobson, who represented the <mask>of New <mask>.
4
LLM Prediction
This is the historical relationship between the stock markets and the over-the counter markets, emphasized Paul Raisler, a partner at Fried, Frank, Harris, Shriver & Jacobson, who represented the City of New York in the case.
2
Keywords Extraction
['Preserving', 'differentiation', 'exchange', 'markets', 'counter', 'markets', 'emphasized', 'Paul', 'Raisler', 'partner', 'Fried', 'Frank', 'Harris', 'Shriver', 'Jacobson', 'represented', 'Banks', 'New', 'York',]
6
Graph Construction
'Preserving', 'differentiation', 'exchange', 'markets', 'counter', 'markets', 'emphasized', 'Paul', 'Raisler', 'partner', 'Fried', 'Frank', 'Harris', 'Shriver', 'Jacobson', 'represented', 'Banks', 'New', 'York', 'This is', 'relationship', 'stock ', 'Frank', 'Harris', City ', 'York in the case'
5
Predicted Keywords
['This is', 'relationship', 'stock ', 'Frank', 'Harris', 'City ', 'York in the case']
8
GCN
7
Similarity Calculation
[('Preserving', 'This is'), ('differentiation', 'relationship), ('exchange', 'stock '), ('Frank', 'Frank'), ('Harris', 'Harris'), ('Banks', 'City '), ('York', 'York in the case')]
9
Binary Classifier
AI
Human
$Z^{(j+1)} = \sigma (D^{-1/2} (A+S)D^{-1/2}Z^{(j)} W)$

Figure 3.1: SGG-ATD detects AI-generated text by constructing a graph per input, where nodes are original and predicted keywords. Edges encode lexical semantics (cosine) and contextual (prediction-based) similarity. A GCN processes the graph for final classification. An example illustrating this process is shown on the right.

constructed graph is processed using a two-layer GCN [43], which propagates and aggregates information across keyword nodes. A global graph representation is then derived and passed to a classifier to determine whether the input text is AI-generated or human-written.

We highlight the novelty and contributions of this framework as follows. (1) *Predictive Masking for Structural Signal:* Unlike prior works, our approach probes contextual predictability by masking semantic keywords and reconstructing them using a pretrained language model, capturing generative patterns often indicative of AI-written text. (2) *Dual Similarity Graph Encoding:* The integration of lexical semantics and contextual similarity into a single graph structure enables more expressive relational modeling. (3) *Graph-Based Reasoning over Prediction-Informed Graphs:* We leverage a Graph Convolutional Network (GCN) over the constructed similarity graph to model higher-order dependencies, supporting robust detection beyond surface-level textual patterns.

3.3.1 Keyword Extraction and Masking

Given an input text, we extract a set of keywords $\mathcal{K} = \{k_1, k_2, \ldots, k_n\}$ using part-of-speech (POS) tagging, focusing on nouns and verbs as they carry core semantic meaning. We randomly select a subset $\mathcal{M} \subset \mathcal{K}$, masking 30% of the keywords by replacing them with `<mask>` tokens:

$$|\mathcal{M}| = \lfloor \alpha n \rfloor, \quad \text{where } \alpha = 0.3 \tag{3.1}$$

This results in a masked version of the input text T_m, which is used to probe contextual predictability in the following stage.

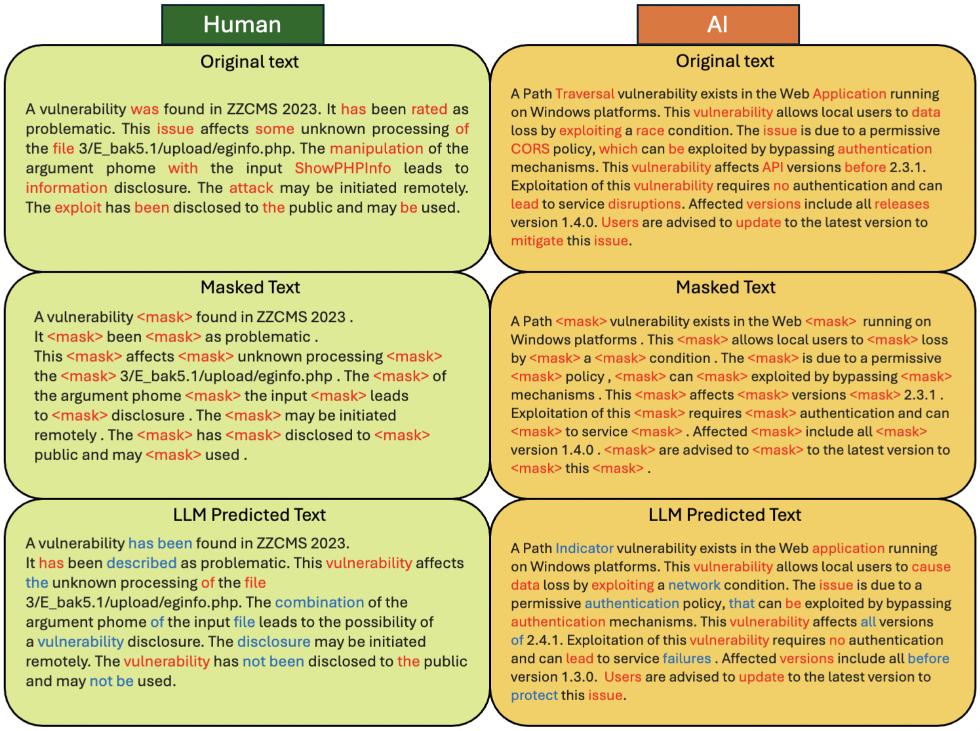
3.3.2 Masked Keyword Prediction

To expose latent structural differences between AI-generated and human-written texts, we employ a prediction step inspired by masked language modeling (MLM). The masked input text is passed to a pretrained ALBERT-base-v2 model [42], which predicts the missing keywords based on surrounding context.

Our hypothesis is that language models demonstrate higher confidence and accuracy in reconstructing masked tokens in AI-generated text, due to its syntactic regularity and high dependency on keyword-level patterns. In contrast, human-written content—being more varied and context-rich—leads to greater prediction uncertainty.

As illustrated in Figure 3.2, this behavioral difference becomes evident when comparing prediction results across both text types. The figure shows that AI-generated texts result in more accurate predictions, while human-written texts often produce more incorrect keywords (incorrect predictions are highlighted in blue), supporting our hypothesis.

The predicted keywords are treated as contextual reconstructions and are later used to construct a graph alongside the original keywords. Formally, given a masked input text T_m, the predicted keywords $\hat{\mathcal{M}}$ are obtained as:



A vulnerability was found in ZZCMS 2023. It has been rated as problematic. This issue affects some unknown processing of the file 3/E_bak5.1/upload/eginfo.php. The manipulation of the argument phome with the input ShowPHPInfo leads to information disclosure. The attack may be initiated remotely. The exploit has been disclosed to the public and may be used.

A vulnerability <mask> found in ZZCMS 2023 . It <mask> been <mask> as problematic . This <mask> affects <mask> unknown processing <mask> the <mask> 3/E_bak5.1/upload/eginfo.php . The <mask> of the argument phome <mask> the input <mask> leads to <mask> disclosure . The <mask> may be initiated remotely . The <mask> has <mask> disclosed to <mask> public and may <mask> used .

A vulnerability has been found in ZZCMS 2023. It has been described as problematic. This vulnerability affects the unknown processing of the file 3/E_bak5.1/upload/eginfo.php. The combination of the argument phome of the input file leads to the possibility of a vulnerability disclosure. The disclosure may be initiated remotely. The vulnerability has not been disclosed to the public and may not be used.

A Path Traversal vulnerability exists in the Web Application running on Windows platforms. This vulnerability allows local users to data loss by exploiting a race condition. The issue is due to a permissive CORS policy, which can be exploited by bypassing authentication mechanisms. This vulnerability affects API versions before 2.3.1. Exploitation of this vulnerability requires no authentication and can lead to service disruptions. Affected versions include all releases version 1.4.0. Users are advised to update to the latest version to mitigate this issue.

A Path <mask> vulnerability exists in the Web <mask> running on Windows platforms . This <mask> allows local users to <mask> loss by <mask> a <mask> condition . The <mask> is due to a permissive <mask> policy , <mask> can <mask> exploited by bypassing <mask> mechanisms . This <mask> affects <mask> versions <mask> 2.3.1 . Exploitation of this <mask> requires <mask> authentication and can <mask> to service <mask> . Affected <mask> include all <mask> version 1.4.0 . <mask> are advised to <mask> to the latest version to <mask> this <mask> .

A Path Indicator vulnerability exists in the Web application running on Windows platforms. This vulnerability allows local users to cause data loss by exploiting a network condition. The issue is due to a permissive authentication policy, that can be exploited by bypassing authentication mechanisms. This vulnerability affects all versions of 2.4.1. Exploitation of this vulnerability requires no authentication and can lead to service failures . Affected versions include all before version 1.3.0. Users are advised to update to the latest version to protect this issue.

Figure 3.2: This illustration highlights the rationale behind our masking strategy, as applied to samples from a vulnerability dataset. In both AI-generated and human-written examples, 30% of the keywords have been masked. The language model predicts these tokens, and the differences in prediction accuracy provide insight into the predictability patterns of each text type. Incorrect predictions (blue tokens) are more frequent in human-written samples, highlighting reduced contextual predictability.

$$\hat{\mathcal{M}} = \text{ALBERT}(T_m) \tag{3.2}$$

To ensure high-quality predictions, we filter out punctuation and malformed outputs (e.g., incomplete tokens, symbols).

3.3.3 Graph Construction with Dual Similarity Encoding

A graph representation of the text is constructed, where nodes represent both original and LLM-predicted keywords. We construct a similarity graph where each node is connected to every other node, and edges are weighted using two key similarity measures:

1. **Lexical Semantic Adjacency Matrix** (A): Captures semantic similarity between

words on subword-level lexical features using FastText [44] embeddings via cosine similarity.

2. **Contextual Similarity Matrix (S):** Encodes contextual alignment between original and predicted keywords based on dot-product similarity.

These two similarity measures are computed independently and reflect distinct aspects of textual structure: lexical semantics and contextual predictability.

The initial lexical semantic adjacency matrix A is computed as:

$$A_{ij} = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|} \tag{3.3}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ are FastText embeddings of words i and j.

The contextual similarity matrix S is given by:

$$S_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j \tag{3.4}$$

where S captures contextual alignment between original and predicted keywords based on masked reconstruction behavior.

To form the final graph structure, we integrate both signals by summing the two matrices:

$$A' = A + S \tag{3.5}$$

The combined adjacency matrix A' is then used as input to the GCN for graph-based reasoning.

3.3.4 Graph-Based Classification via GCN

The constructed similarity graph is processed using a Graph Convolutional Network (GCN), which operates on the enhanced adjacency matrix A' that encodes both lexical semantics and contextual similarity. The GCN propagates information across nodes to refine their embeddings and model higher-order relationships relevant for classification.

Node embeddings are updated layer-wise as follows:

$$Z^{(i+1)} = \sigma\left(D^{-1/2}(A' + I)D^{-1/2}Z^{(i)}W\right) \tag{3.6}$$

where $Z^{(i)}$ is the node embedding at layer i, A' is the modified adjacency matrix, D is the degree matrix, W is a trainable weight matrix, and σ is a non-linear activation function (e.g., ReLU). The initial input $Z^{(0)} = X$ corresponds to the feature matrix composed of FastText embeddings of the original and predicted keywords.

After the final GCN layer, the node representations are aggregated using mean pooling to form a global graph representation, which is passed to a classifier:

$$\hat{y} = \text{softmax}(\text{Classifier}(\text{MeanPool}(Z))) \tag{3.7}$$

Here, $\hat{y}$ is the predicted class label indicating whether the input text is AI-generated or human-written.

3.3.5 Training and Evaluation

The GCN-based classifier is trained using a binary cross-entropy loss function:

$$\mathcal{L} = -\sum_i y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) \tag{3.8}$$

where $y_i \in \{0, 1\}$ is the true label (1 for AI-generated, 0 for human-written), and $\hat{y}_i$ is the predicted probability output from the model.

3.3.6 Summary

We introduce a novel framework that leverages both masked language modeling and graph-based reasoning. By constructing a graph whose adjacency matrix integrates both lexical semantics similarity and prediction-informed contextual relationships, our model captures subtle patterns in text structure and predictability — patterns that are often indicative of machine authorship.

3.4 Evaluations

3.4.1 Datasets

To evaluate our proposed approach, we use four diverse text datasets representing different writing domains and linguistic challenges:

- **News Dataset** – Comprised of journalistic content, featuring a formal tone and fact-based reporting. This dataset was sourced from Verma et al. [15].
- **Creative Writing** – Includes fictional and narrative-driven samples, characterized by varied vocabulary and stylistic choices. This dataset is also based on the collection by Verma et al. [15].
- **Student Essay** – Contains argumentative and academic-style writing, often demonstrating structured reasoning and moderate complexity. The samples are derived from Verma et al. [15].
- **Vulnerability Dataset** – A domain-specific dataset focused on software vulnerability descriptions, which combines technical jargon with concise summaries. We constructed this dataset ourselves: human-written samples were extracted from the National Vulnerability Database (NVD) [45], while AI-generated samples were created using ChatGPT [8] to produce vulnerability descriptions aligned with the style and content of NVD entries.

Table 3.1 provides sample examples from these domains for comparison and Table 3.2 presents the dataset statistics. These datasets are selected to evaluate the model's robustness

Table 3.1: Comparison of AI and Human Samples Across Domains

Datasets	**AI Samples**	**Human Samples**
News	The committee's main task will be to define how the new addresses should be managed and who will legally control them.	The Internet may be overflowing with new technology but crime in cyberspace is still of the old-fashioned variety.
Creative Writing	I shrug. 'It gets old after a while, ya know? Plus, there's not much to do in the same place for over a year.'	'You have finally arrived' He projected into my mind, with the most chilling cold and unhuman voice.
Student Essay	On the other hand, women in many societies may feel pressure to have children due to familial or societal expectations, irrespective of their personal desires. Such societal pressures can contribute to women having children they do not particularly desire, leading to dissatisfaction and regret.	In conclusion why women do or do not have children is a complex process influenced by many factors, and based upon a variety of discourses and opportunities ingrained within society, not simply whether or not a woman likes children.
Vulnerability Dataset	The XML data exchange endpoint does not disable external entity processing, allowing attackers to inject malicious entities. This can lead to unauthorized access to serverside files and even sensitive user data.	Unrestricted Upload of File with Dangerous Type vulnerability in JiangQie Free Mini Program allows Upload a Web Shell to a Web Server. This issue affects JiangQie Free Mini Program: from na through 2.5.2.

Table 3.2: Dataset Statistics Across Domains

	News	**Creative**	**Student**	**Vulnerability**
		Writing	**Essay**	**Dataset**
# Dataset Size	479	728	13629	946
# Median Length	45	38	82	30
# Minimum Length	3	2	2	4
# Maximum Length	208	354	291	429

across a broad spectrum of writing styles and domains, including general-purpose news reporting, academic essays, creative narratives, and highly technical software vulnerability descriptions. This diversity ensures that the model is exposed to varying linguistic patterns, domain-specific vocabulary, and stylistic complexity, making it well-suited for detecting AI-generated content in both generic and specialized contexts. For our experiments, each dataset is randomly split into 80% training and 20% testing subsets.

3.4.2 Implementation Details

We implemented our model in PyTorch [46, 47], leveraging the HuggingFace Transformers library and pretrained `ALBERT-Base v2` [42] for masked language modeling. Keyword extraction was performed using NLTK [48], and FastText embeddings were used to represent nodes in the graph. Each input sample was converted into a graph structure informed by lexical semantics and contextual similarity. A two-layer Graph Convolutional Network (GCN) processed the graph, and its output was passed through a fully connected layer for binary classification. The model was trained using binary cross-entropy loss with the Adam optimizer [49], a learning rate of 0.01, and 100 epochs on an NVIDIA GPU.

3.4.3 Baselines

We compare our proposed method against several state-of-the-art AI-generated text detection approaches that employ diverse detection strategies:

- **GPTZero [50]:** It is a commercially available AI tool that analyzes mathematical features such as perplexity to assess whether a given text is likely written by a human or generated by an AI model.
- **DetectGPT [14]:** A zero-shot method that leverages the curvature of the log-probability landscape in the output space of a language model to identify text likely generated by AI.
- **Ghostbuster [15]:** An approach that constructs feature representations using aggregated predictions from multiple small language models, aiming to capture statistical irregularities in AI-generated text.
- **RAIDAR [24]:** A rewriting-based method that evaluates the degree of textual change introduced by language models when rewriting input passages, using edit distance as a discriminative signal.

3.4.4 Main Results

Table 3.3 presents the core results of our model and baseline comparisons across all four datasets using F1 score as the evaluation metric, consistent with prior works [15, 24] where it was the sole reported metric. Among existing models, RAIDAR and Ghostbuster demonstrate strong performance in structured and technical domains like Student Essay and Vulnerability dataset, reaching up to 0.69 and 0.75 respectively. However, our model, which integrates contextual graph modeling with masked keyword reconstruction, achieves the highest F1 scores across all domains — attaining 0.98 on the Vulnerability dataset and 0.85 on Student Essay using a masking ratio of 0.3.

Table 3.3: Performance comparison (F1 Scores) across all datasets

Methods	**News**	**Creative Writing**	**Student Essay**	**Vulnerability Dataset**
GPTZero (2023)	0.43	0.61	0.48	0.66
DetectGPT (2023)	0.41	0.63	0.52	0.72
GhostBuster (2023)	0.59	0.57	0.64	0.75
RAIDAR (2024)	0.63	0.65	0.69	0.84
SGG-ATD (Ours)	**0.79**	**0.72**	**0.85**	**0.98**

Furthermore, our method significantly outperforms all baselines in challenging domains such as Creative Writing and News, where other detectors like GPTZero and DetectGPT struggle due to reliance on shallow statistical cues. The consistent performance of our model across diverse writing styles — facilitated by the use of a 0.3 masking ratio — demonstrates the robustness and generalizability of our graph-augmented detection framework.

3.4.5 Analysis

Effect of LLM Backbone

As shown in Table 3.4, we evaluate the performance of our framework using different backbone language models for predicting masked keywords with a fixed masking ratio of 0.3. ALBERT-Base v2 achieves the best overall balance across domains, particularly in News and Creative Writing, while also maintaining strong performance in Student Essay and Vulnerability dataset. DeBERTa-Base and Roberta perform competitively, achieving near-identical results on the Vulnerability dataset. Even BERT-Base-Uncased yields strong scores, especially in Student Essay. These results indicate that our graph-augmented framework is modular and model-agnostic, capable of leveraging a range of encoder backbones without substantial performance degradation.

Table 3.4: Performance using different LLMs in our model

LLM (Our Model)	**News**	**Creative**	**Student**	**Vulnerability**
		Writing	**Essay**	**Dataset**
BERT-Base-Uncased	0.75	0.66	0.88	0.97
ALBERT-Base v2	0.79	0.72	0.85	0.98
DeBERTa-Base	0.75	0.72	0.85	0.97
Roberta	0.73	0.70	0.86	0.98

Out-of-Distribution (OOD) Generalization

Table 3.5 presents the out-of-distribution (OOD) evaluation results. For the OOD setting, we adopt a *leave-one-domain-out* evaluation strategy to simulate cross-domain generalization. Specifically, the model is trained on a combination of three datasets (e.g., News, Creative Writing, and Student Essay) and tested exclusively on the remaining unseen dataset (e.g., Vulnerability Dataset). These target unseen domains differ significantly in tone, structure, vocabulary, and syntactic variability—making OOD evaluation a strong indicator of real-world robustness. Our model consistently achieves the highest F1 scores in each domain, including substantial improvements in News (0.67 vs. 0.49 and 0.58) and Vulnerability dataset (0.75 vs. 0.62 and 0.66). Notably, it also outperforms RAIDAR and Ghostbuster in more stylistically varied domains like Creative Writing and Student Essay, indicating strong generalization capabilities.

It is important to note that GPTZero and DetectGPT are unsupervised methods. Therefore, their OOD performance remains identical to their in-domain performance, further highlighting the advantage of our supervised graph-based design in adapting to unseen domains. These results collectively suggest that SGG-ATD is more robust to distributional shifts and adaptable across diverse linguistic domains and writing styles.

Table 3.5: Out-of-Distribution (OOD) Evaluation – F1 Scores

Dataset (F1 Scores)	**News**	**Creative**	**Student**	**Vulnerability**
		Writing	**Essay**	**Dataset**
GPTZero	0.43	0.61	0.48	0.66
DetectGPT	0.41	0.63	0.52	0.72
GhostBuster	0.49	0.52	0.50	0.62
RAIDAR	0.58	0.59	0.53	0.66
SGG-ATD	**0.67**	**0.65**	**0.61**	**0.75**

Effect of Masking Ratio

Figure 3.3 illustrates the effect of different masking ratios on our model's performance across the four datasets. We observe that the Vulnerability dataset and Student Essay datasets remain relatively stable across all masking levels, with the Vulnerability dataset consistently achieving F1 scores above 0.97 and peaking at 0.99 for multiple ratios. In contrast, domains like Creative Writing are more sensitive to the masking ratio; performance declines at higher masking levels, dropping from 0.78 at 0.1 to 0.71 at 0.5 and 0.9. The News dataset shows a gradual improvement up to a masking ratio of 0.3, where it reaches its peak F1 score of 0.79, before plateauing or slightly dropping. Based on these trends, we select a masking ratio of 0.3 as the default in our framework. This ratio offers the best balance across all domains—yielding the highest score in News and competitive results in the other three. It avoids the over-masking that degrades performance in more variable, stylistic domains while still providing enough masked context for the model to learn meaningful reconstruction patterns. Overall, a 0.3 masking ratio supports both stability and generalization, making it

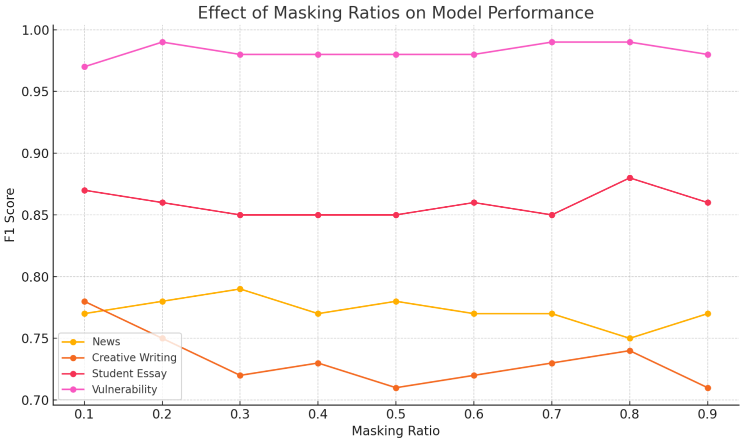


Figure 3.3: Effect of masking ratio on model performance (F1 score) across four datasets. A masking ratio of 0.3 provides a strong balance across domains, achieving the highest score in News and maintaining competitive results elsewhere, while higher ratios degrade performance in stylistically variable datasets like Creative Writing.

an effective setting for our masked keyword-based graph model.

3.5 Conclusions

In this work, we introduced a graph-augmented framework for detecting AI-generated text by leveraging masked keyword reconstruction and contextual relational modeling. By masking a portion of input text and using ALBERT-Base v2 to predict the masked tokens, our approach captures subtle structural and semantic differences between human and AI-written content. We further enriched this signal by constructing a graph of original and predicted keywords, enabling the model to reason over their contextual dependencies. Extensive experiments across four diverse datasets—News, Creative Writing, Student Essay, and Vulnerability dataset—demonstrated that our method consistently outperforms strong baselines such as GPTZero, DetectGPT, Ghostbuster, and RAIDAR.

Additionally, our model exhibited strong generalization to out-of-distribution (OOD) data, and ablation studies on masking ratios revealed that a masking ratio of 0.3 offers the best trade-off across domains. These results validated the robustness, adaptability, and modularity of our approach. Future work may explore extending this framework to multimodal inputs or incorporating dynamic masking strategies to improve adaptability and performance in real-world settings.

Bibliography

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic, "Claude (v1-v3)," https://www.anthropic.com/index/claude, 2023, large language model.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[8] A. Open, "Chatgpt (mar 14 version)[large language model]," 2023.

[9] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[10] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[11] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[12] S. Gehrmann, H. Strobelt, and A. M. Rush, “Gltr: Statistical detection and visualization of generated text,” *arXiv preprint arXiv:1906.04043*, 2019.

[13] 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.

[14] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “Detectgpt: Zero-shot machine-generated text detection using probability curvature,” in *International Conference on Machine Learning.* PMLR, 2023, pp. 24 950–24 962.

[15] V. Verma, E. Fleisig, N. Tomlin, and D. Klein, “Ghostbuster: Detecting text ghostwritten by large language models,” *arXiv preprint arXiv:2305.15047*, 2023.

[16] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps *et al.*, “Release strategies and the social impacts of language models,” *arXiv preprint arXiv:1908.09203*, 2019.

[17] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, “Automatic detection of generated text is easiest when humans are fooled,” *arXiv preprint arXiv:1911.00650*, 2019.

[18] G. Jawahar, M. Abdul-Mageed, and L. V. Lakshmanan, “Automatic detection of machine generated text: A critical survey,” *arXiv preprint arXiv:2011.01314*, 2020.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[20] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[21] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.

[22] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi, “Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text,” *arXiv preprint arXiv:2107.01294*, 2021.

[23] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, and B. Raj, “Gpt-sentinel: Distinguishing human and chatgpt generated content,” *arXiv preprint arXiv:2305.07969*, 2023.

[24] C. Mao, C. Vondrick, H. Wang, and J. Yang, “Raidar: generative ai detection via rewriting,” *arXiv preprint arXiv:2401.12970*, 2024.

[25] A. Valdez and H. Gómez-Adorno, “Text graph neural networks for detecting ai-generated content,” in *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, 2025, pp. 134–139.

[26] A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu, “Eagle: A domain generalization framework for ai-generated text detection,” *arXiv preprint arXiv:2403.15690*, 2024.

[27] A. Bhattacharjee, T. Kumarage, R. Moraffah, and H. Liu, “Conda: Contrastive domain adaptation for ai-generated text detection,” *arXiv preprint arXiv:2309.03992*, 2023.

[28] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *International Conference on Machine Learning.* PMLR, 2023, pp. 17 061–17 084.

[29] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang, “Provable robust watermarking for ai-generated text,” *arXiv preprint arXiv:2306.17439*, 2023.

[30] N. S. Kamaruddin, A. Kamsin, L. Y. Por, and H. Rahman, “A review of text watermarking: theory, methods, and applications,” *IEEE Access*, vol. 6, pp. 8011–8028, 2018.

[31] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models,” *arXiv preprint arXiv:2307.15043*, 2023.

[32] Y. Zhang, Y. Ma, J. Liu, X. Liu, X. Wang, and W. Lu, “Detection vs. anti-detection: Is text generated by ai detectable?” in *International Conference on Information.* Springer, 2024, pp. 209–222.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[34] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *arXiv preprint arXiv:2101.00190*, 2021.

[35] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[36] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” in *The Eleventh International Conference on Learning Representations*, 2022.

[37] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, “Gpt detectors are biased against non-native english writers,” *Patterns*, vol. 4, no. 7, 2023.

[38] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, “How close is chatgpt to human experts? comparison corpus, evaluation, and detection,” *arXiv preprint arXiv:2301.07597*, 2023.

[39] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can ai-generated text be reliably detected?” *arXiv preprint arXiv:2303.11156*, 2023.

[40] M. L. Siddiq, S. H. Majumder, M. R. Mim, S. Jajodia, and J. C. Santos, "An empirical study of code smells in transformer-based code generation techniques," in *2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM)*. IEEE, 2022, pp. 71–82.

[41] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in *International Conference on Acoustics, Speech, and Signal Processing,*. IEEE, 1989, pp. 695–698.

[42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[44] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[45] National Institute of Standards and Technology, "National vulnerability database (nvd)," https://nvd.nist.gov/.

[46] 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.

[47] 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.

[48] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

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

[50] E. Tian, "Gptzero," 2023, aI-generated text detection tool. [Online]. Available: https://gptzero.me

4 Conclusion

This thesis explored two critical and interconnected problem spaces, first, multimodal fake news detection and second, AI-generated text detection, both aimed at safeguarding the integrity of information in the AI era.

In the first part, we presented a novel architecture for fake news detection that incorporates multimodal inputs (text and image) and contextual credibility cues (news domains) using a rotational attention mechanism. By combining BERT embeddings for textual data, VGG-19 features for image content, and domain-level reasoning via Graph Attention Networks (GAT), the model captures rich, cross-modal interactions and improves performance over traditional baselines. This chapter demonstrated that deception in fake news is rarely isolated to one modality and that holistic modeling of visual, textual, and contextual signals is essential.

The second part of this thesis focused on the detection of AI-generated text, an increasingly critical challenge as large language models become more fluent and widely adopted. To address this, we proposed a similarity-guided graph reasoning framework that leverages masked language modeling to predict masked keywords and evaluate semantic coherence. The original and predicted keywords are represented as nodes in a graph, while their pairwise similarities define the edge weights. This graph is then processed through a Graph Convolutional Network (GCN), enabling the model to reason over structural and contextual relationships—resulting in more robust detection, even across diverse and previously unseen generative styles.

Together, the two components of this thesis represent complementary strategies for content authenticity verification. The first focuses on understanding credibility through multimodal fusion and external domain cues, while the second emphasizes semantic consistency and structural coherence in text. Both are designed with the shared goal of building AI systems that can interpret, verify, and protect the quality of information in dynamic and adversarial environments.

This thesis not only introduces novel architectures and reasoning frameworks, but also sets the stage for future research directions such as integrating both tasks into a single real-time detection pipeline, exploring adversarial robustness, and applying these methods to multilingual or cross-platform content. As generative technologies continue to evolve, so too must our tools to identify and mitigate deception—ensuring that innovation in AI is matched by equal progress in AI accountability and content integrity.