Reinforced Adaptive Knowledge Learning for Multimodal Fake News Detection

Litian Zhang[1*], Xiaoming Zhang[1†], Ziyi Zhou[1], Feiran Huang[2], Chaozhuo Li[3*]

[1] Beihang University,
[2]Jinan University,
[3]Beijing University of Posts and Telecommunications
{litianzhang, yolixs, hanlinfeng, ziyizhou}@buaa.edu.cn, huangfr@jnu.edu.cn, lichaozhuobupt@163.com

Abstract

Nowadays, detecting multimodal fake news has emerged as a foremost concern since the widespread dissemination of fake news may incur adverse societal impact. Conventional methods generally focus on capturing the linguistic and visual semantics within the multimodal content, which fall short in effectively distinguishing the heightened level of meticulous fabrications. Recently, external knowledge is introduced to provide valuable background facts as complementary to facilitate news detection. Nevertheless, existing knowledge-enhanced endeavors directly incorporate all knowledge contexts through static entity embeddings, resulting in the potential noisy and content-irrelevant knowledge. Moreover, the integration of knowledge entities makes it intractable to model the sophisticated correlations between multimodal semantics and knowledge entities. In light of these limitations, we propose a novel Adaptive Knowledge-Aware Fake News Detection model, dubbed AKA-Fake. For each news, AKA-Fake learns a compact knowledge subgraph under a reinforcement learning paradigm, which consists of a subset of entities and contextual neighbors in the knowledge graph, restoring the most informative knowledge facts. A novel heterogeneous graph learning module is further proposed to capture the reliable cross-modality correlations via topology refinement and modality-attentive pooling. Our proposal is extensively evaluated over three popular datasets, and experimental results demonstrate the superiority of AKA-Fake.

1 Introduction

The rapid development of social media has facilitated news to be easily accessible, cost-effective, and quickly disseminated in a multimodal form (Khattar et al. 2019). Compared with traditional textual news, multimodal news integrating with images or videos is capable of providing more compelling and enhanced reading experiences (Wang et al. 2018). On the other side of the coin, multimodal contents are more vulnerable to malicious fabrications (e.g., misrepresented or forged images), leading to the rapid dissemination of fake news (Khattar et al. 2019). Therefore, it is urgent to automatically identify the credibility of news to alleviate the risk caused by spreading fake news.

[*]These authors contributed equally.

[†]Xiaoming Zhang is the corresponding author.

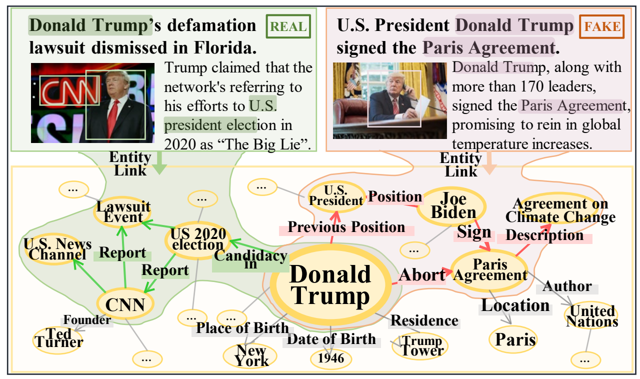


Figure 1: Adaptive knowledge for fake news detection.

Existing detection methods generally focus on capturing the semantics within the news content. Traditional text-based models (Jin et al. 2016; Liu and Wu 2018; Ma, Gao, and Wong 2019) solely rely on modeling the textual linguistic features to make decisions. Multimodal learning models (Chen et al. 2022; Zhu et al. 2022; Tseng et al. 2022) are further introduced to capture both linguistic and visual features, achieving promising advancements by integrating multiple modalities. In order to evade identification from advanced detection models, recent fake news is commonly characterized by a heightened level of meticulous fabrication, rendering it indistinguishable from real news in terms of content patterns (e.g., writing style) (Hu et al. 2021). In response to these emerging challenges, external knowledge is introduced as the complementary, offering reliable factual signals. For example, as shown in the Fig. 1, it is intractable to filter the fake news without the external knowledge "Trump aborted the Pairs Agreement". In a nutshell, both external common knowledge and multimodal news content are indispensable for the successful detection of fake news.

Existing knowledge-enhanced detection methods generally follow the static knowledge-context modeling paradigm (Dun et al. 2021; Hu et al. 2021; Tseng et al. 2022). Entities within the external knowledge graph (KG) (e.g., WikiData) are first encoded into low-dimensional static embeddings via knowledge graph embedding (KGE) (e.g., TransE (Bordes et al. 2013)). For each input news, the entities mentioned in the news are linked to their corresponding entities in the knowledge graph, and the learned entity embeddings are further incorporated as the complementary features. Nonethe-

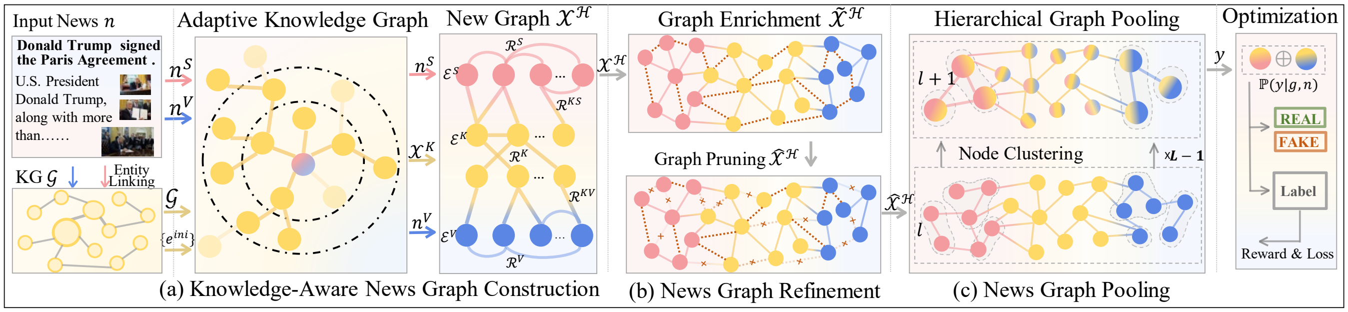


Figure 2: The overview framework of the proposed *AKA-Fake* model.

less, such static knowledge-context modeling methods suffer from the following primary challenges. Firstly, existing KGE methods intrinsically encode all the context neighbors into the representation of the center entity. However, an entity usually has a multitude of neighbors, among which a substantial portion might not be conducive to the fake news detection. For instance, some neighbors like "date/place of birth" and "residence" of the entity "Donald Trump" are uninformative in verifying news in Fig. 1, which may in turn introduce noise potentially. Secondly, the same entity belonging to different news share an identical static embedding learned by KGEs. However, due to the diversity of news, the same entity in various news may necessitate different background knowledge. In Fig. 1, the entity "Donald Trump" in the left news requires the context knowledge about "U.S. 2020 election" marked as the green region, while the identical entity in the right news need context concerning "Pairs Agreement Signing" indicated by the red region. Hence, it becomes imperative to dynamically adapt the context within the knowledge graph to furnish more pertinent information.

To tackle the aforementioned limitations, our motivation lies in generating suitable and adaptive knowledge subgraph, which should be tailored to individual news and pertinent to the objective of detecting fake news. Nevertheless, detecting multimodal fake news with adaptive knowledge is non-trivial. Firstly, the sheer size of the original knowledge graph makes it intractable to adaptively generate a knowledge subgraph that adequately captures the nuances of news content. Secondly, the integration of adapted knowledge entities introduces challenges in effectively modeling the sophisticated correlations between textual semantics, visual objects, and knowledge entities simultaneously. Finally, a multimodal fusion module is essential to facilitate the external knowledge injection and multiple modality aggregation.

In this paper, we propose a novel *Adaptive Knowledge-Aware Fake News Detection* (*AKA-Fake*) model for multimodal news detection. Diverging from the prevailing static knowledge-context modeling paradigm, we propose to generate an adaptive knowledge subgraph based on the news contents and detection annotations, and further incorporate the extracted knowledge into news content via heterogeneous graph learning. First, the adaptive knowledge subgraph is generated by a novel reinforcement learning module, which contributes to conducting an efficient exploration in identifying knowledge paths related to the news background. After that, a heterogeneous news graph is constructed to depict the intricate interactions among various modalities and knowledge entities. In order to alleviate the challenges of sparsity and noisy connections in the news graph, our proposal is integrated with a pioneering graph refinement module to learn the reliable and task-relevant topology. Finally, a hierarchical modality-attentive graph pooling module is proposed to encapsulate the multi-grained cross-modality correlations into the final news representation. *AKA-Fake* is extensively evaluated over three datasets, and the experimental results demonstrate its superiority. Our major contributions are summarized as follows:

- To the best of our knowledge, we are the first to investigate the novel problem of adaptive knowledge learning for multimodal fake news detection.
- A novel *AKA-Fake* model is proposed to learn the adaptive knowledge subgraph under a reinforcement learning paradigm, and integrate knowledge with multimodal content via heterogeneous news graph learning.
- Extensive experiments on real-world datasets reveal that our proposal consistently outperforms SOTA fake news detection baselines.

2 Problem Formulation

***Definition* 2.1** (Multimodal Fake News Detection)**.** *A news is denoted as $n = \{n^S, n^V\}$, where n^S and n^V refer to the textual and visual content within the news, respectively. The textual component, $n^S = \{s_1, s_2, \cdots, s_{L_s}\}$, consists of L_s sentences, while the visual component, $n^{\mathcal{V}} = \{v_1, v_2, \cdots, v_{L_v}\}$, consists of L_v images. The knowledge graph is structured as a collection of semantic triples: $\mathcal{G} = \{(e_h, r, e_t) | e_h, e_t \in \mathcal{E}, r \in R\}$, with $\mathcal{E}$ and R representing the sets of entities and relations, respectively. Our primary goal is to learn a probability distribution that effectively discriminates fabricated news, denoted as $\mathbb{P}(y|n, \mathcal{G})$, where $y = 1$ denotes fake news and 0 stands for real news.*

3 Methodology

The framework of the proposed *AKA-Fake* model is illustrated in Fig. 2, which predicts the label y based on the multimodal news content and the adaptive external knowledge. Mathematically, our proposal is formulated as:

$$\mathbb{P}(y \mid n, \mathcal{G}) = \underbrace{\mathbb{P}(y \mid \hat{\mathcal{X}}^{\mathcal{H}})}_{\substack{\text{Hierarchical}\\ \text{Pooling}}} \underbrace{\mathbb{P}(\hat{\mathcal{X}}^{\mathcal{H}} \mid \mathcal{X}^{\mathcal{H}})}_{\substack{\text{News Graph}\\ \text{Refinement}}} \underbrace{\mathbb{P}(\mathcal{X}^{\mathcal{H}} \mid n, \mathcal{G})}_{\substack{\text{News Graph}\\ \text{Construction}}} \quad (1)$$

where $\mathcal{X}^{\mathcal{H}}$ denotes the initial heterogeneous news graph, and $\hat{\mathcal{X}}^{\mathcal{H}}$ represents the refined news graph. Our proposal consists of three major modules as follows: **Knowledge-Aware News Graph Construction** to generate an adaptive knowledge subgraph through a reinforcement learning framework, **Heterogeneous News Graph Refinement** to learn reliable and task-relevant graph topology $\hat{\mathcal{X}}^{\mathcal{H}}$, and **Modality Attentive Hierarchical Pooling** to learn the final representation of the knowledge-enhanced news.

3.1 Knowledge-Aware News Graph Construction

Here we will introduce the details of adaptive knowledge subgraph generation and news graph construction.

Reinforced Adaptive Knowledge Subgraph Generation. For each input news, we aim to learn its unique adapted knowledge subgraph to provide proper and trustworthy external facts. Considering knowledge graph usually encompasses millions of relations and entities, the straight-forward enumeration strategy is practically unfeasible due to the overwhelming intricacy (Xu et al. 2020). Hence, inspired by previous works (Liu et al. 2021; Park et al. 2022), we formalize the adaptive knowledge subgraph generation as a reinforcement learning problem. The KG is regarded as a versatile framework for preserving an agent's comprehension of news content through the application of reinforcement learning, which contributes to providing explicit reasoning pathways to facilitate decision-making.

• **Multimodal Entity Linkage.** To integrate external knowledge, the first step is to link the entities within the news to their corresponding entities in the KG, dubbed entity linkage. Existing methods (Dun et al. 2021; Hu et al. 2021) primarily emphasize the integration of textual knowledge. Different from existing works, we conduct multimodal entity linkage to comprehensively capture both textual and visual entities. For the textual content n^S, we employ the textual entity linkage tool TAGME [1], which is capable of locating and linking named entities mentioned in the news text into the KG $\mathcal{G}$. For the visual content n^V, the Faster R-CNN (Ren et al. 2016) is applied to detect the important objects such as celebrities and organizations (Qi et al. 2021). Then a pre-trained visual entity linkage model (Hu et al. 2023) is used to link the detected objects into KG based on the similar embedding retrieval. Finally, the linked textual and visual entities are merged as the matched entities $\{e^{ini}\}$.

• **The Markov Decision Process (MDP)**. The reinforcement learning framework is formulated as a classic MDP by a 4-tuple $(\mathcal{S}, \mathcal{A}, \Upsilon, \Gamma)$, where $\mathcal{S}$ denotes the finite states space during exploration, $\mathcal{A}$ represents the action space, $\Upsilon : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ stands for the reward function derived from the environment, and $\Gamma : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ denotes the transition probability function. Fig. 3 presents a toy example of the learning process. The input news is views as the seed node "0", and its connected neighbors are the matched entities $\{e^{ini}\}$. At step t, we aim to identify the most suitable knowledge context neighbors into the constructed graph based on the action selection. The current state s_t and can-

[1]https://sobigdata.d4science.org/web/tagme/

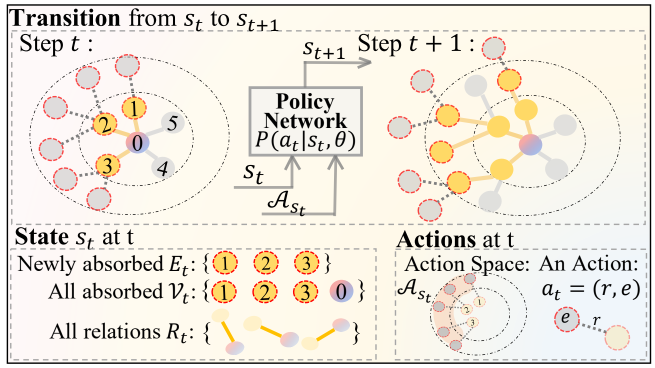


Figure 3: The MDP of reinforced adaptive knowledge subgraph generation.

didate actions are fed into the policy network, which outputs the probability of the action selection, leading to the transition to state $t+1$. The agent will continuously explore the newly knowledge entities until a predetermined condition on the size of absorbed entities set is reached.

At each step, the state $s_t \in \mathcal{S}$ is represented as a 4-tuple $(n, E_t, \mathcal{V}_t, R_t)$, where E_t indicates the newly absorbed entities at step t. $\mathcal{V}_t$ and R_t represent the sets of all absorbed entities and relations up to step t, respectively. The initial state is denoted by $s_0 = (n, \{e_n\}, \emptyset, \{e_n\})$, where e_n represents seed node of the news itself. The action space $\mathcal{A}_t$ for state s_t is defined as the collection of outgoing relationships among entities in E_t, denoted as $\mathcal{A}_t = \{(r,e)|(e_h, r, e) \in \mathcal{G}, e_h \in E_t, e \notin \mathcal{V}_t\}$. The reward function assesses the navigated state's quality, with News Relevant Reward and Terminal Reward serving as distinct measures for environment feedback. Detailed elaboration can be found in Section 3.4. Given a state $s_t = (n, E_t, \mathcal{V}_t, R_t)$ and an action $a_t = \{(r,e)\}$, the transition to the next state is defined as $\mathbb{P}(s_{t+1} = (n, E_{t+1}, \mathcal{V}_{t+1}, R_{t+1})|s_t, a_t) = 1$.

• **Subgraph Generation Policy**. The neural policy network $\pi_\theta(s_t, a_t)$ estimates the conditional probability distribution of actions given state s_t. State s_t is implemented as:

$$\mathbf{s_t} = \left([\boldsymbol{\eta} \| \mathbf{x_t} \| \mathbf{v_t}]\right) \tag{2}$$

where $\|$ denotes the concatenation operation, $\boldsymbol{\eta}$ is the document embedding, $\mathbf{x_t}$ is the vector representation of the subgraph at step t, and $\mathbf{v_t}$ denotes the vector representation of newly added entities at step t. The computation of $\mathbf{x_t}$ involves two steps: the initial refinement of entity embeddings $\mathbf{e_h}$ within its neighborhood $\mathcal{V}_t$, and the aggregation of all entities through an attention network:

$$\alpha_0^\theta(e_h) = \boldsymbol{w}^\theta \mathrm{ELU}(\mathbf{W_1^\theta e_h'}), \mathbf{e_h'} = \mathrm{Tanh}(\mathbf{W_0^\theta}[\mathbf{e_h} \| \sum_{e_t \in \psi(e_h)} \mathbf{e_t}])$$

$$\mathbf{x}_t = \sum_{e_h \in \mathcal{V}_t} \alpha_0^\theta(e_h)\mathbf{e}_h', \quad \alpha_1^\theta(e_h) = \frac{\exp(\alpha_0^\theta(e_h))}{\sum_{b \in E_t} \exp(\alpha_\theta^0(e_b))}$$

where ELU is the exponential linear unit and $\psi(e_h)$ denotes the neighbors of e_h. The embedding of newly added entity $\mathbf{v_t}$ is learned by averaging of the entities within E_t:

$$\mathbf{v_t} = \frac{\sum_{\mathbf{e}_h \in E_t} \mathbf{e}_h}{|E_t|}. \tag{3}$$

Each available action a_i is modeled as $\mathbf{a_i} = (\mathbf{r_i} + \mathbf{e_i})$, denoting as the superposition of two vectors. Then, the policy network calculates the probability of taking an action unit a_i at state s_t as follows:

$$\pi_\theta(s_t, a_i) = \frac{\exp(\mathbf{W_2^\theta}\mathrm{ReLU}(\mathbf{W_3^\theta}[\mathbf{s}_t \| \mathbf{a}_i]))}{\sum_{a_j \in \mathcal{A}_t} \exp(\mathbf{W_2^\theta}\mathrm{ReLU}(\mathbf{W_3^\theta}[\mathbf{s}_t \| \mathbf{a}_j]))}. \quad (4)$$

To accelerate the generation process, we let the agent absorb multiple outgoing relations at one action a_t (receptive field), instead of only absorbing one relation (Jiang et al. 2019).

Heterogeneous News Graph Construction. The incorporation of external knowledge enhances the intricate and complex relationships among diverse modalities. To capture such sophisticated correlations, the input news is represented by a heterogeneous graph $\mathcal{X}^{\mathcal{H}}$ following previous work (Ragesh et al. 2021). $\mathcal{X}^{\mathcal{H}}$ contains three types of nodes: knowledge nodes $\mathcal{E}^K$ contains the absorbed knowledge entities, sentence nodes $\mathcal{E}^S$ in which each node denoting a sentence within the news, and visual nodes $\mathcal{E}^V$ consists of visual objects (Qi et al. 2021; Zhang, Zhang, and Pan 2022) from images. Five types of edges are constructed as follows: $\mathcal{R}^{KS}$ links sentence nodes to the knowledge entities, $\mathcal{R}^{KV}$ connects the visual nodes and the linked entities, $\mathcal{R}^{KK}$ denotes the original connections within the external knowledge graph among the absorbed entities, $\mathcal{R}^{SS}$ connects sentences by words co-occuring in fixed-size sliding window (Piao et al. 2022; Yan et al. 2023), and $\mathcal{R}^{VV}$ links the visual objects belonging to the same image.

3.2 Heterogeneous News Graph Refinement

Existing works directly learn the embedding based on the constructed news graph (Vaibhav and Hovy 2019; Li et al. 2022). However, the news graph is generated by the heuristic assumptions, which may suffer from the challenges of sparsity, noisy and task-irrelevant edges (Zheng et al. 2020; Li et al. 2017). Here we introduce a graph refinement module to learn reliable and task-relevant topology structures.

News Graph Enrichment. The manually defined graph construction strategies may lead to the long-tail nodes with scarce connections, which further plague the model performance (Pang et al. 2022; Zhao et al. 2022). Here we introduce a relation-wise graph enrichment approach to enhance five types of relations concerning heterogeneous graphs (Zhao et al. 2021). The relation $\mathcal{R}^{\mathcal{KS}}$ is adopted as an example to exemplify the graph enrichment process.

Initially, a semantic graph is generated to capture the semantic similarities between textual sentences and knowledge entities. The sentence node $e^S \in \mathcal{E}^S$ and knowledge node $e^{\mathcal{K}} \in \mathcal{E}^K$ are embedded into latent vectors $h^{\mathcal{S}}$ and $h^{\mathcal{K}}$, respectively. Subsequently, metric learning is employed to learn the semantic graph $\mathbf{G}^{\mathcal{KS}} \in \mathbb{R}^{|\mathcal{E}^{\mathcal{K}}| \times |\mathcal{E}^{\mathcal{S}}|}$ as:

$$\mathbf{G}^{\mathcal{KS}}[i,j] = \begin{cases} \Omega^{\mathcal{KS}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) & \Omega^{\mathcal{KS}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) \geq \epsilon \\ 0 & \Omega^{\mathcal{KS}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) < \epsilon \end{cases}$$

where ϵ is a hyper-parameter to control sparsity. Function $\Omega^{\mathcal{KS}}$ is the z-head weighted cosine similarity:

$$\Omega^{\mathcal{KS}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) = \frac{1}{z}\sum_{i=1}^{z} \cos(\boldsymbol{w}^{\mathcal{KS}} \odot \mathbf{h_i^{\mathcal{K}}}, \boldsymbol{w}^{\mathcal{KS}} \odot \mathbf{h_j^{\mathcal{S}}}) \quad (5)$$

in which $\odot$ is the Hadamard product and $\mathbf{w_0}$ is a learnable vector indicating the importance of different dimensions.

Furthermore, in light of the motivation that nodes with similar semantics tend to share similar neighbors, we introduce two propagation graphs to model the correlations between node semantics and graph typologies. Here we take the knowledge node as an example. First, the similarity matrix $\mathbf{M}^{\mathcal{K}} \in \mathbb{R}^{|\mathcal{E}^{\mathcal{K}}| \times |\mathcal{E}^{\mathcal{K}}|}$ between knowledge is calculated as:

$$\mathbf{M}^{\mathcal{K}}[i,j] = \begin{cases} \Omega^{\mathcal{K}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{K}}}) & \Omega^{\mathcal{K}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) \geq \epsilon \\ 0 & \Omega^{\mathcal{K}}(\mathbf{h_i^{\mathcal{K}}}, \mathbf{h_j^{\mathcal{S}}}) < \epsilon \end{cases} \quad (6)$$

where $\Omega^{\mathcal{K}}$ is same to Formula (5). The knowledge propagation matrix $\mathbf{P}^{\mathcal{K}} \in \mathbb{R}^{|\mathcal{E}^{\mathcal{K}}| \times |\mathcal{E}^{\mathcal{S}}|}$ is obtained as follows:

$$\mathbf{P}^{\mathcal{K}} = \mathbf{M}^{\mathcal{K}}\mathbf{A}^{\mathcal{KS}} \quad (7)$$

in which $\mathbf{A}^{\mathcal{KS}} \in \mathbb{R}^{|\mathcal{E}^{\mathcal{K}}| \times |\mathcal{E}^{\mathcal{S}}|}$ denotes to the original adjacency matrix with relation $\mathcal{R}^{\mathcal{KS}}$. The sentence propagation matrix $\mathbf{P}^{\mathcal{S}} \in \mathbb{R}^{|\mathcal{E}^{\mathcal{K}}| \times |\mathcal{E}^{\mathcal{S}}|}$ can be obtained similarly.

Finally, the original graph in $\mathbf{A}^{\mathcal{KS}}$, similarity graph $\mathbf{S}^{\mathcal{KS}}$ and two propagation graphs $\mathbf{P}^{\mathcal{K}}$ and $\mathbf{P}^{\mathcal{S}}$ are combined together as enriched graph for $\mathcal{R}^{\mathcal{KS}}$:

$$\widetilde{\mathbf{A}}^{\mathcal{KS}} = \Gamma(\mathbf{A}^{\mathcal{KS}}, \mathbf{G}^{\mathcal{KS}}, \mathbf{P}^{\mathcal{K}}, \mathbf{P}^{\mathcal{S}}) \quad (8)$$

where Γ is the combination function implemented as the plus operator to ensure the efficiency. $\widetilde{\mathbf{A}}^{\mathcal{KS}}$ denotes the enriched graph topology of relation $\mathcal{R}^{\mathcal{KS}}$.

News Graph Pruning. The augmented edges learned by the graph enrichment contribute to alleviating the challenge of sparsity, while may in turn introduce potential noise. Here, we propose a meta-path based graph pruning method to remove the task-irrelevant edges.

Given a specific meta-path Φ and center node e, graph pruning aims to remove noisy meta-path based neighbors from $\mathcal{N}^{\Phi}(e)$. $\mathcal{N}^{\Phi}(e)$ is the set of nodes connected with e through meta-path Φ. Here we still take the meta-path $\mathcal{KS}$ as an example. We propose to sample the L-neighborhood subgraph $\hat{\mathbf{A}}^{\mathcal{KS}}$ from the enriched graph $\widetilde{\mathbf{A}}^{\mathcal{KS}}$. Subgraph $\hat{\mathbf{A}}^{\mathcal{KS}}$ shares the same node set to $\widetilde{\mathbf{A}}^{\mathcal{KS}}$, and each node $e^{\mathcal{K}}$ in $\hat{\mathbf{A}}^{\mathcal{KS}}$ can keep no more than L edges. The edge importance score of the node $e_j^{\mathcal{S}} \in \mathcal{N}^{\Phi^{\mathcal{KS}}}(e_i^{\mathcal{K}})$ to the node $e_i^{\mathcal{K}}$ is defined as:

$$d_{i,j}^{\mathcal{KS}} = \mathrm{sigmoid}\left(\mathrm{FFN}\left(\left[\mathbf{h_i^{\mathcal{K}}} \| \mathbf{h_j^{\mathcal{S}}} \| \widetilde{A}^{\mathcal{KS}}[i,j]\right]\right)\right). \quad (9)$$

The scores of different edges are achieved as $\mathbf{D}_i^{\mathcal{KS}} = \{d_{i,1}^{\mathcal{KS}}, d_{i,2}^{\mathcal{KS}}, \ldots, d_{i,N}^{\mathcal{KS}}\}$ in which N is number of meta-path neighbors. Given the scores between center nodes with neighborhoods, we sample Top-K edges with highest scores. Furthermore, we define the select result $\mathbf{I} = [I_1, I_2, \cdots, I_D] \in \{0,1\}^{N \times K}$ as one-hot N dimensional indicator vectors. Then the Top-K with sorted scores is equivalent to the following linear program:

$$\arg\max_{\mathbf{I} \in \mathbf{C}} \langle \mathbf{I}, \mathbf{D}\mathbf{1}^\top \rangle \quad (10)$$

where $\mathbf{D}$ is simplified form of $\mathbf{D}_i^{\mathcal{KS}}$. $\mathbf{D}\mathbf{1}^\top \in \mathbb{R}^{N\times K}$ is constructed by replicating scores K times. $\langle\rangle$ denotes flattening the matrices and taking dot product. The constraint set C is:

$$C = \Big\{\mathbf{I} \in \mathbb{R}^{N\times K} : \mathbf{I}_{n,l} \geq 0, \mathbf{1}^\top \mathbf{I} = 1, \mathbf{I}\mathbf{1} \leq 1\ , \sum_{i\in[N]} i\mathbf{I}_{i,l} < \sum_{j\in[N]} j\mathbf{I}_{j,l'}\ \forall l < l'\Big\}\,. \tag{11}$$

The first condition encourages that each of the columns in $\mathbf{I}$ has a total weight of one. The last condition results in sorting the nodes. This operation is non-differentiable because both Top-K and one-hot operations are non-differentiable.

To learn the parameters using back propagation, the perturbed maximum method (Berthet et al. 2020) is introduced to generate discrete samples by adding uniform Gaussian noise $\mathbf{Z}$ and perturb the input $\mathbf{D}$ in the train stage:

$$\mathbf{I}_\delta = \mathbb{E}_{\mathbb{Z}}\left[\arg\max_{\mathbf{I}\in\mathcal{C}} \langle \mathbf{I}, \mathbf{D}\mathbf{1}^\top + \delta\mathbf{Z}\rangle\right] \tag{12}$$

where δ is a hyper-parameter to control the input weight of noisy. During training, we linearly decay δ to zero. At $\delta = 0$, no noise is added and the differentiable Top-K operation is numerically identical to hard Top-K. Finally, edge represented by $\mathbf{I}_\delta$ is added into the pruned adjacency matrix $\hat{\mathbf{A}}^{\mathcal{KS}}$. The pruned heterogeneous graph $\hat{\mathcal{X}}^{\mathcal{H}}$ can be achieved by applying similar process over different meta-paths.

3.3 Modality Attentive Hierarchical Pooling

To learn the final news representation, a modality attentive graph pooling module is proposed to capture the multimodal content hierarchically. *AKA-Fake* first learns the cluster assignment matrix and processes graph coarsening hierarchically (Zhang et al. 2023). Then the adaptive knowledge graph is aggregated with other modalities . Here we take textual graph $\mathcal{X}^S$ as an example. Inspired by (Ying et al. 2018; Chen et al. 2020), the l-th cluster assignment matrix is as follows:

$$\mathbf{C}_l^S = \mathrm{softmax}(\mathrm{GAT}_l^{pool}(\hat{\mathbf{F}}_l^S, \mathbf{A}_l^S)) \tag{13}$$

where $\hat{\mathbf{F}}_l^S \in \mathbb{R}^{|\mathcal{E}_l^S|\times d}$ is the node feature embedding, $\mathbf{A}_l^S \in \{0,1\}^{|\mathcal{E}_l^S|\times|\mathcal{E}_l^S|}$ is the adjacency matrix, $\mathbf{C}_l^S \in \mathbb{R}^{|\mathcal{E}_l^S|\times|\mathcal{E}_{l+1}^S|}$ is the cluster assignment matrix. , which $|\mathcal{E}_{l+1}^S|$ is the cluster number at layer $l+1$. $|\mathcal{E}_{l+1}^S| = \rho|\mathcal{E}_l^S|$ is , where ρ is a hyper-parameter to control cluster number. GAT is the graph attention network (Veličković et al. 2018). Then the graph coarsening is formulated as:

$$\begin{aligned}\mathbf{F}_{l+1}^S &= (\mathbf{C}_l^S)^T \cdot \mathrm{GAT}_l^{emb}(\hat{\mathbf{F}}_l^S, \mathbf{A}_l^S)\\ \mathbf{A}_{l+1}^S &= (\mathbf{C}_l^S)^T \mathbf{A}_{l+}^S \mathbf{C}_l^S\end{aligned} \tag{14}$$

where $\mathbf{F}_{l+1}^S \in \mathbb{R}^{|\mathcal{E}_{l+1}^S|\times d}$ is the pooling textual feature, $\mathbf{A}_{l+1}^S \in \{0,1\}^{|\mathcal{E}_{l+1}^S|\times|\mathcal{E}_{l+1}^S|}$ is the adjacency matrix. After the graph coarsening in every layer, we apply meta-path based aggregation to incorporate the adaptive knowledge:

$$\hat{\mathbf{F}}_{l+1}^S = \text{Meta-Agg}(\mathbf{F}_{l+1}^S, \mathbf{A}_{l+1}^S, \{\Phi\}). \tag{15}$$

Meta-Agg is the meta-path based node aggregation (Wang et al. 2019). At the penultimate layer $L-1$, the assignment matrix $\mathbf{C}_{L-1}$ is a vector of 1. Thus, all nodes at final layer L are assigned to a single cluster, which is the final textual embedding $\hat{\mathbf{h}}^{\mathbf{V}} = \mathbf{F}_L^S \in \mathbb{R}^{1\times d}$. After achieving the final visual embedding $\hat{\mathbf{h}}^{\mathbf{S}} \in \mathbb{R}^{1\times d}$ in the similar manner, these two embedding are concatenated and fed into a prediction layer to make the final decision:

$$\mathbb{P}(y \mid \hat{\mathcal{X}}^{\mathcal{H}}) = \mathrm{sigmod}(\mathbf{W_P}[\hat{\mathbf{h}}^{\mathbf{S}} \| \hat{\mathbf{h}}^{\mathbf{V}}] + \mathbf{b}). \tag{16}$$

3.4 The Optimization Framework

In this section, we introduce the multi-task optimization framework to train the *AKA-Fake* model.

Reinforcement Learning Reward. The reinforced adaptive knowledge subgraph generation process contains News Relevant Reward and Terminal Reward.

- **News Relevant Reward**. The agent is encouraged to select entities which have relevant background information to the current news n. To this end, we build a reverse index to track all the news that mention an entity e in the title: $\Lambda(e) = \{n_1, n_2, \cdots, n_k | n_i\ contains\ e\}$. The reward is:

$$\mathcal{R}^B = \sum_{e_i\in a_m} \mathcal{R}'(n, e_i)\mathcal{R}'(n, e_i) = cosine(\mathbf{v}_n, \frac{\sum_{n_i\in\Lambda(e_i)} \mathbf{v}_i}{|\Lambda(e)|}). \tag{17}$$

- **Terminal Reward.** We consider only the terminal reward at the last time point T, when the adaptive knowledge subgraph has already been generated. The reward is :

$$\mathcal{R}^T = \mathbb{P}(y \mid \mathcal{X}^K, n)\mathcal{R}'(y). \tag{18}$$

$\mathcal{R}'(y) = \mathbb{I}(y = y')$, which is 1 (or 0) if y is (or is not) equal to the ground-truth label y'. $\mathcal{R}(t)$ is a compound reward:

$$\mathcal{R}(t) = \lambda \cdot \mathcal{R}^B(t) + (1-\lambda) \cdot \mathcal{R}^T(t) \tag{19}$$

where $\lambda \in [0, 1]$ is a weighting coefficient.

Actor-Critic Optimization. The critic evaluates an action by estimating the action value function $Q(s, a)$ in the MDP environment. For the critic network, we use a similar structure to the policy network (Eq. 4):

$$Q_\phi(s_t, a_i) = \mathbf{W_Q^0}\mathrm{ReLU}\big(\mathbf{W_Q^1}[\mathbf{s_t}\|\mathbf{a_i}]\big). \tag{20}$$

The critic network is trained with the Temporal Difference (TD) method (Sutton 1988). It first calculates a target q_t according to the Bellman (Bellman 2013) equation:

$$q_t = \mathcal{R}(t) + \mathbb{E}_{a\sim\pi_\theta}\left[\gamma \cdot \mathcal{Q}_\phi(s_{t+1}, a)\right] \tag{21}$$

where γ is a decay factor. The critic network is learned by minimizing the TD error:

$$\mathcal{L}_{critic} = (Q_\phi(s_t, a_t) - q_t)^2 \tag{22}$$

The actor aims to maximize the expected payoff of each step w.r.t. the given Q-function:

$$J_{actor}(\theta) = \mathbb{E}_{a\sim\pi_\theta}\left[Q_\phi(s_t, a)\right]. \tag{23}$$

We use the policy gradient method (Sutton et al. 2000) to optimize the actor network. The gradients of $J(\theta) w.r.t. \theta$ are calculated as follows for each sampled trajectory:

$$\nabla_\theta J_{actor}(\theta) \simeq Q_\phi(s_t, a_t) \cdot \nabla_\theta log\pi_\theta(s_t, a_t) \tag{24}$$

The fake news detection cross-entropy loss is:

$$\mathcal{L}_{fnd} = y^* \log(\mathbb{P}(y \mid \mathcal{G}, n) + (1-y^*)\log(1-\mathbb{P}(y \mid \mathcal{G}, n)).$$

Category	Method	PolitiFact				GossipCop				Pheme			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Unimodal Method	VGG-19	0.458	0.492	0.473	0.482	0.443	0.478	0.462	0.470	0.537	0.522	0.579	0.549
	TextCNN	0.608	0.621	0.623	0.622	0.733	0.698	0.703	0.701	0.583	0.787	0.549	0.647
	BERT	0.781	0.766	0.878	0.818	0.759	0.787	0.764	0.775	0.768	0.704	0.778	0.739
Multimodal Method	EANN	0.804	0.808	0.794	0.801	0.796	0.812	0.765	0.788	0.771	0.714	0.701	0.707
	MVAE	0.812	0.803	0.835	0.819	0.782	0.802	0.751	0.776	0.776	0.735	0.723	0.728
	SAFE	0.872	0.883	0.897	0.890	0.830	0.843	0.894	0.868	0.815	0.799	0.795	0.797
	SpotFake	0.770	0.753	0.795	0.773	0.812	0.807	0.822	0.814	0.823	0.868	0.863	0.865
	CAFE	0.848	0.847	0.860	0.853	0.832	0.804	0.903	0.851	0.830	0.876	0.855	0.865
Knowledge Enhanced	KAN	0.859	0.869	0.850	0.860	0.777	0.776	0.770	0.773	0.783	0.759	0.744	0.751
	CompNet	0.830	0.869	0.860	0.865	0.825	0.869	0.824	0.846	0.825	0.888	0.856	0.872
	KAHAN	0.869	0.883	0.912	0.897	0.827	0.864	0.917	0.890	0.824	0.793	0.842	0.817
Proposed	**AKA-Fake**	**0.919**	**0.930**	**0.930**	**0.930**	**0.856**	**0.880**	**0.941**	**0.909**	**0.858**	**0.918**	**0.877**	**0.897**
	Imp.(%)	*+4.7*	*+4.7*	*+1.8*	*+3.3*	*+2.4*	*+1.1*	*+2.4*	*+1.9*	*+2.8*	*+3.0*	*+1.4*	*+2.7*

Table 1: Overall performance of different detection models. Best results are in bold and second best results are underlined.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on three popular datasets: PolitiFact, GossipCop (Shu et al. 2020), and Pheme (Qi et al. 2019), which contain 495, 15,707 and 2,099 news, respectively. The dataset is randomly split into training, validation, and testing sets with a ratio of 7:1:2.

Baselines. Three types of SOTA baselines are adopted, including **the unimodal models**: VGG-19 (Simonyan and Zisserman 2014), TextCNN (Chen 2015) and BERT (Devlin et al. 2018); **the multimodal models**: EANN (Wang et al. 2018), MVAE (Khattar et al. 2019), SAFE (Zhou, Wu, and Zafarani 2020), SpotFake (Singhal et al. 2019) and CAFE(Chen et al. 2022); and **the KG enhanced models**: KAN (Dun et al. 2021), CompNet (Hu et al. 2021) and KAHAN (Tseng et al. 2022).

Implementation Details. We use CLIP (Radford et al. 2021) as the pre-trained model for text and image embedding. The respective field size is set to [5, 3, 2]. We employ Adam (Kingma and Ba 2014) as the optimizer. The batch size is set as 64. The initial learning rate is set to $2e^{-4}$. The embedding size for entities, relations, and document is 1024.

4.2 Quantitative Evaluation

Each detection model is executed five times, and the average results are reported in Table 1. One can clearly see that the text-based methods (TextCNN and BERT) surpass visual-based ones (VCG-19), demonstrating that textual semantics are more important. Multimodal methods generally outperform unimodal ones, indicating the distinct and valuable contributions provided by various modalities. After introducing external knowledge, detection performance is further improved, revealing the pivotal role of external knowledge. Our proposal consistently outperforms all baselines over three datasets, achieving +4.7% on PolitiFact, +2.4% on GossipCop, and +2.8% on Pheme in terms of accuracy. The superiority of *AKA-Fake* verifies the effectiveness of adaptive knowledge incorporation. In summary, *AKA-Fake* excels in its capacity to intricately capture multimodal content and effectively assimilate reliable external knowledge, culminating in superior overall performance.

4.3 Ablation Study

In this section, ablation studies are conducted to investigate the importance of different modules. Table 2 demonstrates the results of different ablation models.

• **Multimodal Input**. We aim to study the influence of different modalities. “w/o Image”, “w/o KG” and “w/o Img. & KG” stand for the input without image, KG, and both, respectively. One can achieve the following observations. (1) Knowledge is more important than the images due to the more significant performance decline of “w/o KG”. (2) Both of these modalities are crucial as model performance presents a drop after removing a single modality.

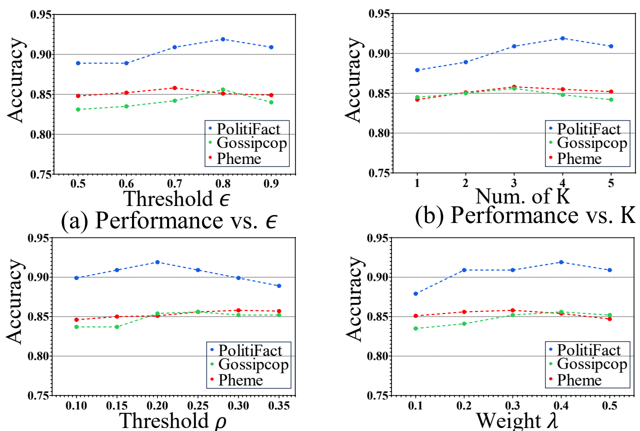
• **Adaptive Knowledge Subgraph Generation**. Here we focus on investigating the effectiveness of adaptive knowledge learning. “Static 1-order” and “Static 2-order” denote the incorporation of all knowledge context neighbors within 1 or 2 orders. “Node Similarity” is designed to select top-10 context neighbors based on the semantic similarities. (1) After replacing the adaptive subgraph generation with conventional static neighbors, model performance present significant decline, verifying that static KGE embeddings may introduce noises. (2) Compared with the manually defined selection strategy “Node Similarity”, our proposal achieves better performance, revealing that the learning of knowledge subgraph is capable of capturing the task-relevant and content-related external knowledge.

• **Graph Refinement**. Here we focus on the influence of Graph Refinement. “w/o Enrich.”, “w/o Pruning”, “w/o Refinement” are implemented without Graph Enrichment, Graph Pruning, and both modules. (1) Without the whole refinement module, performance declines significantly due to the sparsity and noises in the news graph constructed following heuristic assumptions. (2) Both modules are crucial to achieving desirable detection performance as both “w/o Enrich.” and “w/o Pruning” only achieve degraded results.

• **Modality Attentive Hierarchical Pooling**. “Mean Pooling” and “Max Pooling” are applied to replace Hierarchical Pooling. The absence of hierarchical pooling results in noticeable performance degradation. This observation substantiates that the proposed pooling module effectively integrates adaptive knowledge into the news graph,

Categoty	Method	PolitiFact				GossipCop				Pheme			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Proposed	AKA-Fake	0.919	0.930	0.930	0.930	0.856	0.880	0.941	0.909	0.858	0.918	0.877	0.897
Modality Input	w/o Img. &KG	0.848	0.855	0.895	0.875	0.806	0.837	0.855	0.846	0.802	0.859	0.841	0.850
	w/o image	0.909	0.930	0.912	0.920	0.844	0.877	0.928	0.902	0.844	0.917	0.857	0.886
	w/o KG	0.869	0.919	0.850	0.883	0.819	0.849	0.906	0.877	0.817	0.865	0.877	0.871
Adpative KG Generation	Static 1-order	0.879	0.883	0.912	0.897	0.818	0.842	0.885	0.863	0.826	0.892	0.857	0.874
	Static 2-order	0.889	0.893	0.912	0.902	0.825	0.854	0.891	0.872	0.830	0.887	0.870	0.879
	Node Similarity	0.879	0.872	0.880	0.876	0.817	0.839	0.870	0.855	0.828	0.872	0.862	0.867
News Graph Refinement	w/o Enrich	0.879	0.855	0.842	0.848	0.811	0.842	0.869	0.855	0.815	0.868	0.823	0.845
	w/o Pruning	0.869	0.919	0.842	0.879	0.839	0.875	0.929	0.901	0.835	0.899	0.864	0.881
	w/o Refine	0.869	0.872	0.830	0.850	0.823	0.861	0.908	0.884	0.831	0.882	0.843	0.862
Graph Pooling	Mean Pooling	0.909	0.919	0.912	0.916	0.843	0.872	0.922	0.896	0.839	0.894	0.877	0.885
	Max Pooling	0.899	0.930	0.895	0.912	0.845	0.875	0.920	0.897	0.849	0.912	0.870	0.891

Table 2: Ablation studies of *AKA-Fake* on three datasets.



(a) Performance vs. ϵ (b) Performance vs. K (c) Performance vs. ρ (d) Performance vs. λ

Figure 4: Hyper-parameter sensitivity analysis.

leading to a discernible enhancement in model performance.

4.4 Parameter Sensitivity Analysis

Here we conduct hyper-parameter sensitivity analysis on the weights of four parameters: sparsity controller ϵ in Eq. (6), number of sampled nodes K in Eq. (10), cluster number controller ρ in Eq. (13) and balance parameter λ in Eq. (19). In Fig. 4 (a), with the increase of ϵ, model performance first increases and then significantly drops. A larger ϵ leads to a sparser graph with limited topology information, while noise may be introduced if ϵ is too small. Fig. 4 (b) shows that with the increase of sampling number K, model performance first improves and then drops. This demonstrates that appropriate meta-path based neighbors would be sufficient to provide enough topology information. Fig 4 (c) shows that model performance presents a similar trend with the increase of ρ. The smaller ρ leads to less clusters with potential information missing, while a larger ρ may burden the training process. In Fig. 4 (d), with the increase of the News Relevant Reward, performance increase and drop slowly. Thus, λ should be carefully tuned to achieve desirable performance.

4.5 Case Study

Fig. 5 presents two cases along with the learned knowledge subgraphs denoted by different colors. For the identical entity "Beyonce", different knowledge contexts are selected according to the news content. For instance, the entity "Beyonce" in the first news concerns "Spouse", while her family background is more significant in the second news. Hence, the knowledge subgraph generated by *AKA-Fake* is highly relevant to the news content background, providing adaptive and reliable knowledge to enhance detection performance.

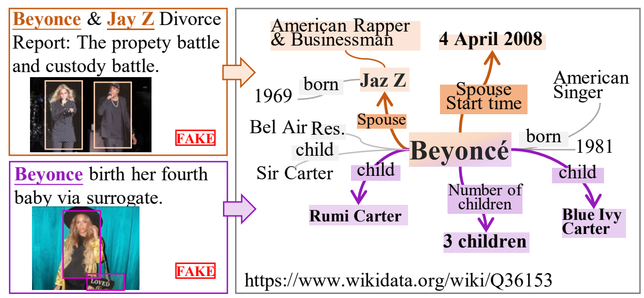


Figure 5: Case studies on GossipCop dataset.

5 Related Work

We divide fake news detection methods into two categories. **Content-based**: many methods focus on the single modality content features for detection (Gupta et al. 2013; Ma et al. 2016; Ma, Gao, and Wong 2017; Jin et al. 2016; Liu and Wu 2018; Ma, Gao, and Wong 2019; Qi et al. 2019). Recent works have also considered multimodal content in news (Jin et al. 2017; Wang et al. 2018; Khattar et al. 2019; Zhou, Wu, and Zafarani 2020; Wu et al. 2021; Singhal et al. 2022; Chen et al. 2022). **Knowledge-enhanced**: some studies use the linked entity in KG as external knowledge to supplement news content (Hu et al. 2021; Dun et al. 2021; Qi et al. 2021; Qian et al. 2021; Tseng et al. 2022).

6 Conclusion

In this paper, we propose a novel Adaptive Knowledge-Aware Fake News Detection model (*AKA-Fake*). We generate an adaptive knowledge subgraph based on the news contents, and further incorporate it into news via heterogeneous graph learning. Experimental results evaluated on three popular datasets demonstrate the superiority of our proposal.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62272025 and No.U22B2021), and the Fund of the State Key Laboratory of Software Development Environment.

References

Bellman, R. 2013. Dynamic programming, courier corporation. *New York, NY*, 707.

Berthet, Q.; Blondel, M.; Teboul, O.; Cuturi, M.; Vert, J.-P.; and Bach, F. 2020. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33: 9508–9519.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chen, C.; Qian, S.; Fang, Q.; and Xu, C. 2020. Hapgn: Hierarchical attentive pooling graph network for point cloud segmentation. *IEEE Transactions on Multimedia*, 23: 2335–2346.

Chen, Y. 2015. *Convolutional neural network for sentence classification*. Master's thesis, University of Waterloo.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dun, Y.; Tu, K.; Chen, C.; Hou, C.; and Yuan, X. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 81–89.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736.

Hu, H.; Luan, Y.; Chen, Y.; Khandelwal, U.; Joshi, M.; Lee, K.; Toutanova, K.; and Chang, M.-W. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 754–763.

Jiang, J.; Dun, C.; Huang, T.; and Lu, Z. 2019. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*.

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

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3): 598–608.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, 163–179. Springer.

Li, R.; Zhao, J.; Li, C.; He, D.; Wang, Y.; Liu, Y.; Sun, H.; Wang, S.; Deng, W.; Shen, Y.; et al. 2022. House: Knowledge graph embedding with householder parameterization. In *International Conference on Machine Learning*, 13209–13224. PMLR.

Liu, D.; Lian, J.; Liu, Z.; Wang, X.; Sun, G.; and Xie, X. 2021. Reinforced anchor knowledge graph generation for news recommendation reasoning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1055–1065.

Liu, Y.; and Wu, Y.-F. 2018. 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*, volume 32.

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

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Ma, J.; Gao, W.; and Wong, K.-F. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*, 3049–3055.

Pang, B.; Li, C.; Liu, Y.; Lian, J.; Zhao, J.; Sun, H.; Deng, W.; Xie, X.; and Zhang, Q. 2022. Improving relevance modeling via heterogeneous behavior graph learning in bing ads. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3713–3721.

Park, S.-J.; Chae, D.-K.; Bae, H.-K.; Park, S.; and Kim, S.-W. 2022. Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 784–793.

Piao, Y.; Lee, S.; Lee, D.; and Kim, S. 2022. Sparse structure learning via graph neural networks for inductive document classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11165–11173.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1212–1220.

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

Qian, S.; Hu, J.; Fang, Q.; and Xu, C. 2021. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3): 1–23.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Ragesh, R.; Sellamanickam, S.; Iyer, A.; Bairi, R.; and Lingam, V. 2021. Hetegcn: heterogeneous graph convolutional networks for text classification. In *Proceedings of the 14th ACM international conference on web search and data mining*, 860–868.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.

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

Singhal, S.; Pandey, T.; Mrig, S.; Shah, R. R.; and Kumaraguru, P. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In *Companion Proceedings of the Web Conference 2022*, 726–734.

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

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3: 9–44.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient method for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12: 1057–1063.

Tseng, Y.-W.; Yang, H.-K.; Wang, W.-Y.; and Peng, W.-C. 2022. KAHAN: knowledge-aware hierarchical attention network for fake news detection on social media. In *Companion Proceedings of the Web Conference 2022*, 868–875.

Vaibhav, R. M. A.; and Hovy, E. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. *EMNLP-IJCNLP 2019*, 134.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

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

Xu, K.; Song, L.; Feng, Y.; Song, Y.; and Yu, D. 2020. Coordinated reasoning for cross-lingual knowledge graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 9354–9361.

Yan, H.; Li, C.; Long, R.; Yan, C.; Zhao, J.; Zhuang, W.; Yin, J.; Zhang, P.; Han, W.; Sun, H.; et al. 2023. A Comprehensive Study on Text-attributed Graphs: Benchmarking and Rethinking. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11676–11684.

Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J. B.; Zhang, Y.; Xie, X.; Wang, H.; and Kim, S. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 168–176.

Zhao, J.; Qu, M.; Li, C.; Yan, H.; Liu, Q.; Li, R.; Xie, X.; and Tang, J. 2022. Learning on large-scale text-attributed graphs via variational inference. *arXiv preprint arXiv:2210.14709*.

Zhao, J.; Wang, X.; Shi, C.; Hu, B.; Song, G.; and Ye, Y. 2021. Heterogeneous graph structure learning for graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4697–4705.

Zheng, C.; Zong, B.; Cheng, W.; Song, D.; Ni, J.; Yu, W.; Chen, H.; and Wang, W. 2020. Robust graph representation learning via neural sparsification. In *International Conference on Machine Learning*, 11458–11468. PMLR.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Zhu, Y.; Sheng, Q.; Cao, J.; Nan, Q.; Shu, K.; Wu, M.; Wang, J.; and Zhuang, F. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.