Graph with Sequence: Broad-Range Semantic Modeling for Fake News Detection

Anonymous Author(s)

ABSTRACT

The rapid proliferation of fake news on social media threatens social stability, creating an urgent demand for more effective detection methods. While many promising approaches have emerged, most rely on content analysis with limited semantic depth, leading to suboptimal comprehension of news content. To address this limitation, capturing broader-range semantics is essential yet challenging, as it introduces two primary types of noise: fully connecting sentences in news graphs often adds unnecessary structural noise, while highly similar but authenticity-irrelevant sentences introduce feature noise, complicating the detection process. To tackle these issues, we propose BREAK, a broad-range semantics model for fake news detection that leverages a fully connected graph to capture comprehensive semantics while employing dual denoising modules to minimize both structural and feature noise. The semantic structure denoising module balances the graph's connectivity by iteratively refining it between two bounds: a sequence-based structure as a lower bound and a fully connected graph as the upper bound. This refinement uncovers label-relevant semantic interrelations structures. Meanwhile, the semantic feature denoising module reduces noise from similar semantics by diversifying representations, aligning distinct outputs from the denoised graph and sequence encoders using KL-divergence to achieve feature diversification in high-dimensional space. The two modules are jointly optimized in a bi-level framework, enhancing the integration of denoised semantics into a comprehensive representation for detection. Extensive experiments across four datasets demonstrate that BREAK significantly outperforms existing methods in identifying fake news. Code is available at https://anonymous.4open.science/r/BREAK.

KEYWORDS

Fake news detection, broad-range semantics, bi-level optimization, graph neural network

ACM Reference Format:
Anonymous Author(s). 2024. Graph with Sequence: Broad-Range Semantic Modeling for Fake News Detection. In . ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The Internet's rapid growth has elevated social media platforms like X (formerly known as Twitter) to vital sources of daily information.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

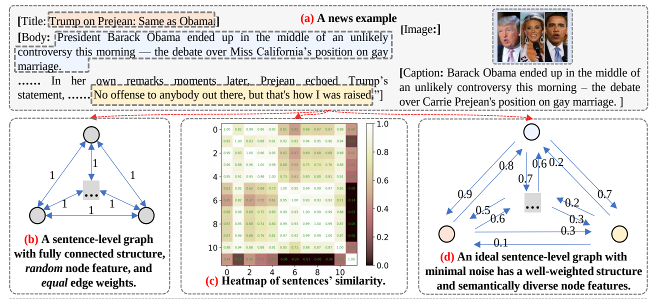


Figure 1: Comparison of traditional and denoised sentence-level graph representations. (a) A news example with both textual and visual information. (b) A traditional sentence-level graph that introduces structural noise and overlooks the semantics of individual sentences. (c) A heatmap showing high sentence similarity in news content. (d) The denoised sentence-level graph proposed in this work, with a well-learned structure and diverse features that reduce noise and preserve semantics.

However, social media provides users freedom to create and disseminate information, making it an "ideal" environment for fake news dissemination [12]. The proliferation of fake news poses great threats to social safety [1] and public health [4][23], thereby necessitating fake news detection as a prominent and urgent research topic.

Throughout the development of fake news detection, news content, as the core component of news, has consistently played a pivotal role. Therefore, fully modeling news content is essential and offers distinct advantages, particularly for the early detection when only news content is available [5]. Moreover, it is beneficial for methods incorporating external data (e.g., comments [24]), as it provides a more comprehensive representation of news content, facilitating better integration with external data [26]. This has been supported by extensive research, including the use of sequence-based and graph-based models [7]. Sequence-based approaches focus on extracting sentimental, semantic, and sometimes cross-modal features from news content, particularly within news texts [38] [30][13]. These texts typically encompass several hundred words, often reiterating key facts and establishing a logical structure replete with clear contextual connections [2]. However, sequence-based methods, better suited for shorter texts, struggle to capture these extended, broader-range connections, overlooking essential contextual nuances [33].

In contrast, graph-based methods strive to encapsulate broad-range semantics by depicting news text as graphs at either the word level [32] or sentence level [36]. Typically, these methods link nodes via sliding windows or fully connected structures for in-depth contextual analysis. However, they often fail to thoroughly unearth the intricate semantics embedded within complex news articles. This shortfall is particularly evident in their limited capacity to fully capture broad-range semantics and the semantic interrelations between nodes. For instance, word-level incorporate lower noise but it also overlooks semantic interrelations outside the sliding window [2][26], and generates an overwhelming number of nodes for lengthy news articles, resulting in substantial computational complexity. A few limited studies have explored sentence-level fully connected graphs to address these shortcomings [28]. These studies employ *random initialization* for node features and then extract patterns of interrelation between sentences to detect fake news.

Although studies based on fully connected graphs successfully capture broader-range semantics, they encounter a significant obstacle: while a fully connected graph structure provides comprehensive coverage, it tends to introduce noise in both the graph structure and the feature representations, such as irrelevant connections and highly similar node features, as illustrated in Figures 1(b) and 1(c). This raises the question: *how can we effectively model broader range semantics while minimizing the introduction of noise?* Handling this noise is essential but comes with two critical challenges. Firstly, the semantic interrelations within a news article can be extremely complex. For example, news articles with 10 sentences could have 90 ($2 \times \Sigma_{k=1}^{9} k$) directed edges, each with different weights, making it hard to eliminate noise and capture key interrelations. Secondly, the semantics of news sentences often exhibit high similarity, hindering the differentiation of key sentences, as depicted in Figure 1(c). These challenges of denoising significantly impede the comprehensive extraction of news content and undermine detection performance, highlighting the need for innovative solutions.

To address the challenges of fake news detection, we introduce BREAK—a model designed to denoise and integrate broad-range semantics for fake news detection. BREAK effectively tackles the noise in both structural and feature semantics, which often obstruct accurate detection, by harmonizing graph-based and sequence-based representations.

BREAK addresses structural noise through an edge weight inference mechanism in its semantic structure denoising module. It first models news content as a fully connected bidirectional graph to capture all potential semantic interrelations. Then, it refines this structure by integrating sequential semantics as a lower bound, treating the fully connected graph as an upper bound and progressively narrowing the graph based on semantically relevant connections. This dynamic process strengthens important links while filtering out irrelevant ones, thereby reducing structural noise. To handle feature noise, BREAK employs a semantic feature diversification mechanism in the semantic feature denoising module. A graph encoder captures broad-range semantics across the entire news article, while a sequence encoder preserves more accurate sequential semantics. By aligning these two distinct representations using KL-divergence, BREAK ensures feature diversification and mitigates the noise from redundant or irrelevant sentence-level semantics.

Through this combined denoising process, BREAK produces a comprehensive and refined semantic representation, significantly enhancing its effectiveness in detecting fake news. The innovation of BREAK lies in its dual denoising approach, which balances capturing broad-range semantics with maintaining sentence-order precision, ultimately leading to more accurate detection. Our contributions in this research are threefold:

- Broad-Range Semantic Modeling for Fake News Detection. We propose BREAK, a novel model using a fully connected graph to capture comprehensive semantics, with dual denoising modules reducing structural and feature noise for more refined, accurate semantic representation.
- Dual Denoising Mechanisms. We design two denoising modules: the semantic structure denoising module, which iteratively refines graph connectivity between a sequence-based lower bound and a fully connected upper bound to uncover label-relevant semantics, and the semantic feature denoising module, which reduces noise from similar semantics by aligning outputs from graph and sequence encoders using KL-divergence.
- Extensive Experimental Validation. We perform extensive experiments on four distinct datasets, showing that BREAK significantly outperforms existing methods in fake news detection. The results show the effectiveness of the bi-level framework in integrating denoised semantics and improving detection performance.

2 RELATED WORK

Existing fake news detection methods focus on news content or additionally incorporating supplementary information (e.g., comments [24], propagation networks [41] and user interactions [37]). However, methods leveraging external information often neglect in-depth exploration of the news content, potentially oversimplifying complex news articles into mere nodes within social networks [31]. This simplification is particularly problematic during the initial stages of news dissemination, where the extra data can be both insufficient and irrelevant [5][30]. Meanwhile, external data like comments may be susceptible to massive noise [24]. Hence, we focus on a foundational approach to fake news detection, i.e., exploring news content.

Sequence-based methods mainly treat news text as sequence data, utilizing sequential models like recurrent neural networks (RNNs) or pre-trained language models (PLM) to capture news features. [34] constructs a hierarchical RNN network to capture the incongruent between the body text and news title. [12] devises a novel prompt paradigm to fully extract the semantics of news by a PLM. Moreover, [13] models multi-modal news through a progressive fusion network, which fuses cross-modal features from various levels. However, these methods fail to explicitly model the broad-range semantics hidden in complex news content [33].

Therefore, some graph-based methods depict news content as a word-level or sentence-level graph, which explicitly represents semantics between words or sentences by an edge. [33] and [6] construct a word-level content graph based on a 2- and 3-size sliding window, respectively, while [32] constructs a multi-hop (3- or 4-hop depending on the dataset) word-level graph to capture syntax relations. However, as outlined in Section 1, these graphs constructed with sliding windows may neglect certain semantics. Hence, several methods construct a sentence-level graph for news content [28] [10] [36]. These methods randomly initialize node features and detect fake news through the interaction patterns between sentences, which ignore the semantics of news sentences.

In the field of text classification in natural language processing (NLP), many counterparts represent text as graphs from word or sentence perspectives. However, they are also suffering from the two challenges we mentioned before. For instance, sliding window

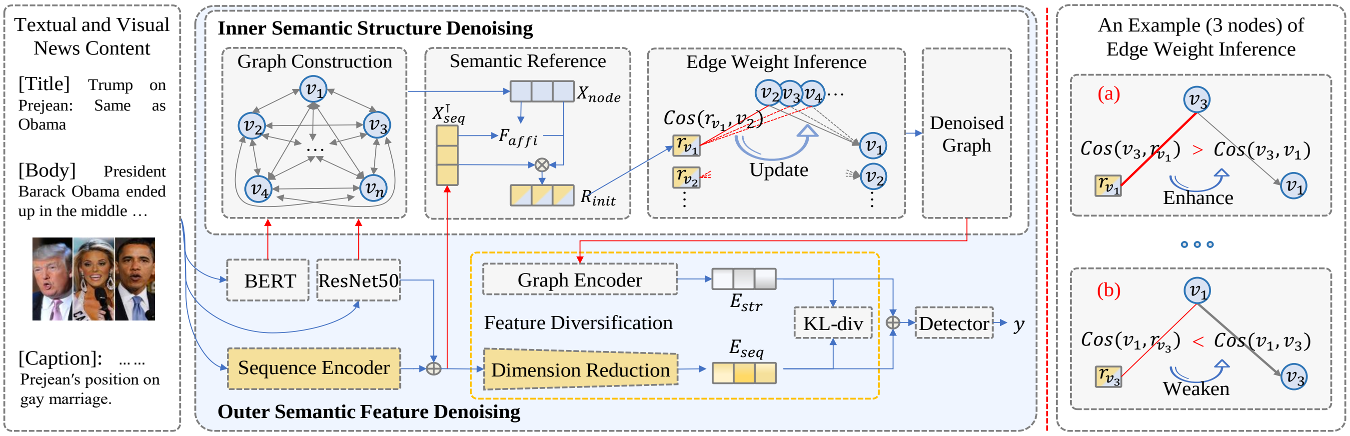


Figure 2: Overview of BREAK with Edge Weight Inference Example. The left part illustrates the overall process of BREAK, where X_{seq} and E_{seq} denote the sequence features that used as the lower bound of semantics, R_{init} indicates the reference semantics that integrate structural and sequential semantics. X_{node} and F_{affi} separately denote the node features and affinity matrix, and E_{str} represents the structural features of the denoised graph. The right part depicts an example of edge weight inference with three nodes.

graphs [39] or sequence graphs (connected according to the original order of words or sentences) [2] neglect some broad-range semantics. Meanwhile, sentence-level graphs [17] or master-node-enhanced graphs [18] (introduce a master node to connect to every other node) lead to densely connected and over-smoothing node features [2].

In Summary, BREAK is different from previous works in two aspects: (1) We represent news content as a more general graph structure, and any extension of existing works will confront the challenges we encountered, such as a large-sized sliding window also introducing vast noise. (2) We devise an effective network to denoise the irrelevant information and extract key sentences while only relying on news content.

3 APPROACH

BREAK aims to detect fake news solely based on news content by capturing broad-range semantics. It involves two modules under the bi-level optimization paradigm: (1) inner semantic structure denoising and (2) outer semantic feature denoising, as illustrated in Figure 2. Specifically, our fully connected graph naturally captures full-range semantics by modeling all potential structural information, while inevitably introducing both structural and feature noise. Therefore, the inner module utilizes the sequential semantics provided by the sequence model as a lower bound to help mitigate structural noise. Subsequently, by aligning the representations of the graph and sequence encoders in the high-dimensional space, the outer module diversifies semantic features and extracts broad-range semantics based on the denoised graph structure.

3.1 Problem Definition

Without loss of generality, we leverage both textual (T) and visual (I) news content for fake news detection, as they are commonly found on social media and are increasingly popular among readers [13] [5]. The textual content encompasses the news title, body text, caption, and others, providing a comprehensive representation of the news textual content. We formally define a piece of fake news with these elements in two forms: (1) The sequential semantics E_{seq} obtained by a sequence encoder and image vectorization tool. (2) The structural semantics E_{str} acquired by a graph encoder. Our final objective is to train a detector f to classify news as real ($y=0$) or fake ($y=1$) based on E_{seq} and E_{str}, i.e., $f(E_{seq}, E_{str}) \rightarrow y \in \{0, 1\}$.

3.2 The Overview of BREAK

BREAK aims to achieve the following three objectives: (1) uniting the structural and sequential semantics to facilitate the narrowing of the lower and upper bounds and obtain a denoised graph structure; (2) diversifying semantic features to realize feature denoising while integrating broad-range semantics; and (3) enabling effective fake news detection.

The structure denoising process depends on the semantic features, and a well-learned semantic feature is determined by the denoised graph structure, indicating a close and mutual influence between them. Therefore, we introduce bi-level optimization to iteratively optimize these two denoising processes. For a piece of news, BREAK first acquires its sequential semantics from the sequence encoder and then represents the news content as a bidirectional fully connected graph to capture all potential semantics. Accordingly, with the support of sequential semantics, BREAK learns a denoised graph structure. In the outer module, a graph encoder is introduced to capture the denoised semantic interrelations and integrate them with the sequential semantics as the broad-range semantics. Ultimately, BREAK detects the authenticity of this news based on these broad-range semantics.

Formally, we represent BREAK as a detect function $f_{\phi,\theta}(\cdot)$, where ϕ indicates the parameters in the inner optimization process, and θ denotes other parameters of BREAK except ϕ. The overall framework can be outlined as follows:

$$\theta^* = \underset{\theta}{argmin}\ \mathcal{L}_{outer}(f_{\phi^*,\theta}(\cdot), Y), \tag{1}$$

$$s.t.\ \phi^* = \underset{\phi}{argmin}\ \mathcal{L}_{inner}(f_{\phi,\theta}(\cdot), Y), \tag{2}$$

where θ^* and ϕ^* are the optimal parameters, $Y \in \{0, 1\}$ is the ground-truth labels of news, and $\mathcal{L}_{inner}$ and $\mathcal{L}_{outer}$ depict the loss functions used in the inner and outer levels, respectively.

3.3 Inner Semantic Structure Denoising

In this process, we model news sentences and images as a bidirectional fully connected graph to cover all potential semantics. Subsequently, by introducing short-rang sequential semantics from the sequence model as the lower bound, we devise a structural and sequential semantics integration mechanism to integrate semantics from both the graph and sequence. These semantics then prompt the edge weight inference mechanism to denoise the graph structure.

Graph Construction. We split all the news textual (T) and visual (I) information in a news article into sentences and images, and then we vectorize them by pre-trained BERT [14] and ResNet50 [8]:

$$X_T = BERT_g(T), \quad X_I = ResNet50(I), \tag{3}$$

where $BERT_g$ and $ResNet50$ are utilized as learnable vectorization tools to learn the sentence ($X_T \in \mathbb{R}^{M\times d}$) and image ($X_I \in \mathbb{R}^{P\times d}$) node features, respectively. Moreover, M and P separately indicate the number of sentences and images. Subsequently, we construct a sentence-level fully connected graph, where each sentence or image corresponds to a node in the graph. Furthermore, following the previous works [11] [36], every two nodes in the graph are connected by a bi-directed edge to reflect the forward and backward context information, as depicted in Figure 2. Therefore, the graph can be represented as $G = (V, E)$ with the adjacency matrix A, edge weight matrix W_e, and feature matrix $X_{node} = (X_T, X_I)$, where both A and W_e are all-one matrix, $X_{node} \in \mathbb{R}^{N\times d}$, $N = M + P$ indicates the node number of G, and d depicts the hidden dimension of node feature.

Structural and Sequential Semantics Integration. We aim to alleviate the noise in G, i.e., enhancing the key semantics and weakening others. Specifically, to reduce the computational complexity and avoid optimizing the graph structure and features simultaneously (since it is more difficult than optimizing either aspect alone), we only modify the graph structure by adjusting the edge weights, which can also guide the aggregation process of graph features.

A fully connected graph produces massive noise, disrupting the original order of sentences, and node features are highly similar (as shown in Figure 1(d)). Due to the aforementioned issues, it is challenging to adjust edge weights within the graph aggregation mechanism without manipulating node features. However, the sequential features preserve the original sentences' order and provide clean (well-learned) sequential semantics, which can be used as a lower-bound representation of a fully connected graph to prompt the denoise process.

In particular, we employ another pre-trained BERT to acquire sentence sequential features by $S_T = BERT_s(T) \in \mathbb{R}^{M\times d}$. Note that we employ **two** independent pre-trained BERTs for different purposes. $BERT_g$ is utilized as a node feature generator, where sequential semantics are disrupted by a fully connected structure. In contrast, $BERT_s$ preserves the natural order of news sentences during the optimization, which provides **cleaner** sequential semantics than a fully connected graph. At last, we concatenate S_T with X_I to obtain the sequential content features as $X_{seq} = S_T \oplus X_I$.

Based on $X_{seq} \in \mathbb{R}^{N\times d}$, we assign an affinity matrix $F \in \mathbb{R}^{N\times N}$ to weigh the affinity between the node and sequential content features. Accordingly, node features $R \in \mathbb{R}^{N\times d}$ that integrate the node and sequential content features (i.e., the reference semantics) can be calculated as follows:

$$F = tanh(X_{node} W_F X_{seq}^{\top}), \tag{4}$$

$$R = tanh(X_{node} W_{node} + F X_{seq} W_{seq}), \tag{5}$$

where W_F, W_{node}, and $W_{seq} \in \mathbb{R}^{d\times d}$ are the weight parameters. R is effective *since it considers broader-range semantics from node and sequential content simultaneously*.

Edge Weight Inference. As depicted in the right part of Figure 2, the edge weight inference process is conducted based on the initially integrated semantics. Taking the nodes v_1 and v_3 as an example, we first calculate the affinity between them and their corresponding sequential content features by Eq. (4): $f_{v_1}, f_{v_3} \in F$. Subsequently, their initially integrated node features $r_{v_1}, r_{v_3} \in R$ are acquired by employing Eq. (5).

In Figure 2(a), v_1 is the target node that we need to update all its incoming edge weights, and v_3 is the source node that we need to justify how much information we should aggregate from it to v_1. Specifically, we treat r_{v_1} as the golden representation of v_1 since it comprehensively integrates broad-range semantics by Eq. (5). Therefore, to guide the aggregation process in the outer optimization and make v_1 **closer** to r_{v_1}, we adjust the incoming edge weights of v_1 according to the cosine similarity between r_{v_1} and other nodes except v_1 (i.e., v_2 and v_3 in this toy example).

This stems from the finding that when $Cos(v_3, r_{v_1}) > Cos(v_3, v_1)$, i.e., v_3 is more similar with r_{v_1} than with v_1, which indicates v_3 owns more golden information in r_{v_1}. Thus, we leverage $Cos(v_3, r_{v_1})$ to enhance the edge weight of $e_{v_3,v_1} \in E$ and thereby ***aggregating more information*** from v_3, and finally make v_1 closer to r_{v_1} in this indirect way. Similarly, this deduction is applicable to the cases of $Cos(v_3, r_{v_1}) < Cos(v_3, v_1)$ and $Cos(v_3, r_{v_1}) = Cos(v_3, v_1)$. In contrast, in Figure 2(b), v_3 is the target node, v_1 is the source node, and all the deductions are the same as in Figure 2(a).

To date, by employing the structural and sequential semantic integration and the edge weight inference for **all** nodes in the graph, the inner structure denoising module has obtained the denoised graph. Meanwhile, the corresponding adjacent matrix and node features of $\hat{G}$ are represented as $\hat{A} = \hat{W_e} \cdot A$ and $\hat{X}_{node} = X_{node}$, respectively, where $\hat{W_e} \in \mathbb{R}^{N\times N}$ indicates the learned edge weights.

3.4 Outer Semantic Feature Denoising

Relying on the denoised graph from the inner optimization process, we can learn distinguishable node features by integrating the sequential semantics (X_{seq}) and the graph representations (enriched with broad-range semantics). This integration ultimately benefits the edge weight inferring process since it helps in learning better sequential semantics and node features.

Semantic Feature Diversification. In detail, we first obtain the graph representation E_{str} by feeding the denoised graph $\hat{G}$ into a graph encoder (which is concretized as a graph convolutional network in this work). Accordingly, a two-layer multi-layer perceptron (MLP) is utilized to align the dimension of short-range semantics

X_{seq} with structural features E_{str}.

$$E_{str} = GCN(\hat{G}), \quad E_{seq} = MLP_{seq}(X_{seq}), \tag{6}$$

where E_{seq} denotes the representation of X_{seq} after dimensionality reduction. $E_{str}, E_{seq} \in \mathbb{R}^{N\times h}$ and h indicates the dimension of hidden state. Specially, E_{str} is the general representation of the initially integrated braod range semantics as it aggregates all the node features.

Consequently, the E_{str} is denoised and contains richer broad-range semantics than in E_{seq}, while the E_{seq} preserves better short-range semantics since it is strictly generated according to the sentences' natural order. Therefore, to assist in diversifying semantic features and then capturing better broad-range semantics, we bring E_{str} and E_{seq} closer in the high-dimensional space to **achieve alignment** by employing the KL divergence as the loss function $\mathcal{L}_{KL}$:

$$\mathcal{L}_{KL} = KL(E_{str}, E_{seq}). \tag{7}$$

Fake News Detection. At last, we concatenate E_{str} and E_{seq} as the **eventual broad-range semantics** to perform fake news detection by a two-layer MLP (MLP_{pred}), and we leverage the binary cross-entropy function as the classify loss function as follows:

$$\hat{Y} = MLP_{pred}(E_{str} \oplus E_{seq}), \tag{8}$$

$$\mathcal{L}_{cls} = -\sum_{i=1}^{Z}[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)], \tag{9}$$

where $\hat{Y}$ represents the predicted labels of news and $\hat{y}_i \in \hat{Y}$. $\mathcal{L}_{cls}$ represents the binary cross-entropy function, Z is the number of news articles, and $y_i \in Y$ indicates the ground-truth label of the news Z_i.

3.5 Model Training

Inner-Level Training. The final objective of the inner level is uniting the semantics from both the graph and sequence encoders, thereby learning a denoised graph structure. We optimize it by maximizing the mutual information between the news representation $E = E_{str} \oplus E_{seq}$ and the new label Y to extract label-related sentences. Specifically, this mutual information can be described as:

$$I(E;Y) = \sum_{E,Y} \mathbb{P}(E,Y) log \frac{\mathbb{P}(E,Y)}{\mathbb{P}(E), \mathbb{P}(Y)}. \tag{10}$$

Since the $\mathbb{P}(Y|E) = \mathbb{P}(E,Y)/\mathbb{P}(E)$ is intractable, inspired by the previous work [16], we introduce a parameterized variational approximation $\mathbb{P}_{\phi}(Y|E)$ for $\mathbb{P}(Y|E)$. Accordingly, we convert the maximization problem into a minimization problem with an upper bound as follows, and the detailed derivation can be found in the Appendix:

$$\phi^* = \underset{\phi}{argmin} \; -I(E;Y), \tag{11}$$

$$-I(E;Y) \leq -\mathbb{E}[\log\mathbb{P}_{\phi}(Y|E)] - H(Y), \tag{12}$$

where the ϕ^* is the optimal parameters of the denoise process. Meanwhile, since $H(Y)$ is the entropy of news labels Y (a constant), we only optimize $-\mathbb{E}[\log\mathbb{P}_{\phi}(Y|E)]$ in the inner level training process. Furthermore, $\mathbb{P}_{\phi}(Y|E)$ is essentially work as a predictor with the parameter ϕ. Specifically, for a piece of news Z_i, BREAK obtains its representation $E_i \in E$ by Eq. (6) and performs a prediction: $f_{\phi,\theta}(Z_i) \to Y_i$. Therefore, $\mathbb{P}_{\phi}(Y_i|E_i)$ is equivalent to the possibility of news Z_i be predicted as Y_i by $f_{\phi,\theta}(Z_i)$. Note that θ is frozen during the inner-level training process, and only the change on ϕ leads to a change on $\mathbb{P}_{\phi}(Y_i|E_i)$. Ultimately, by expanding $-\mathbb{E}[\log\mathbb{P}_{\phi}(Y|E)]$, the loss function of the inner level can be written as a standard cross entropy loss:

$$\mathcal{L}_{inner} = -\sum_{i=1}^{Z}[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)]. \tag{13}$$

Outer-Level Training. In the outer-level training step, our final goal is to diversify semantic features and learn a comprehensive news representation, which captures the broad-range semantics hidden in the news content. Therefore, we combine Eq. (7) and Eq. (9) as the final loss function of the outer level.

$$\mathcal{L}_{outer} = \mathcal{L}_{cls} + \beta\mathcal{L}_{KL}, \tag{14}$$

where β is a hyperparameter used to decide how close the E_{str} and E_{seq} should be.

Table 1: The statistics of the datasets.

Dataset	GossipCop	PolitiFact	Snopes	PolitiFact-S
# fake news	2,466	329	3,177	1,701
# real news	9,270	331	1,164	1,867
# images	11,736	298	0	0

4 EXPERIMENTS

This section evaluates the effectiveness of BREAK on four datasets and aims to answer the following research questions.

RQ1: How does BREAK compare to baselines in fake news detection? **RQ2**: How well does BREAK generalize in scenarios when clear evidence is available? **RQ3**: Is every component of BREAK essential for fake news detection? **RQ4**: How does hyperparameter β influence BREAK's performance? **RQ5**: Can BREAK effectively denoise the fully connected graph and capture crucial semantics?

4.1 Datasets

We perform fake news detection on four real-world datasets as follows to assess the detection performance of BREAK.

Content-Only Datasets. FakeNewsNet [25] comprises two datasets: GossipCop and PolitiFact. Both datasets are sourced from fact-checking websites and labeled as either fake or real. Each news article includes titles and body text, and some also incorporate images.

Clear Evidence Available Datasets. To assess the generalization ability of BREAK in the scenario where clear evidence (one type of external data) is available, we additionally employ two datasets containing evidence: Snopes [21] and PolitiFact [29] (renamed as PolitiFact-S for differentiation). Each entry in these datasets consists of a brief news and its corresponding evidence, and the preprocessed version from [33] is utilized for a fair comparison. The detailed statistics of these datasets are in Table 1.

4.2 Baselines

We compare BREAK with eleven representative and recent baselines as follows:

Table 2: Performance comparison of different methods on GossipCop and PolitiFact datasets, with the best performances in bold and the runners-up underlined. GET is not involved because it cannot work without evidence.

Dataset	Methods	Acc.	Prec.	Rec.	F1
GossipCop	BERT	0.831	0.821	0.831	0.792
	GAT+2 Attn Heads	0.846	0.835	0.846	0.840
	SAFE	0.819	0.817	0.819	0.818
	HMCAN	0.836	0.825	0.836	0.825
	CAFE	0.814	0.824	0.814	0.819
	MRML	0.817	0.838	0.817	0.821
	ALGM	0.829	0.812	0.829	0.811
	CSFND	0.835	0.849	0.835	0.847
	ChatGLM2-6B	0.856	0.847	0.856	0.850
	LLaMA2-7B	<u>0.866</u>	<u>0.858</u>	<u>0.860</u>	<u>0.860</u>
	BREAK	**0.882**	**0.876**	**0.882**	**0.871**
	Improve(%)	1.848	2.098	1.848	1.279
PolitiFact	BERT	0.791	0.794	0.791	0.790
	GAT+2 Attn Heads	0.892	0.893	0.893	0.892
	SAFE	0.853	0.814	0.875	0.843
	HMCAN	<u>0.924</u>	<u>0.927</u>	<u>0.924</u>	0.923
	CAFE	0.791	0.800	0.791	0.793
	MRML	0.817	0.838	0.817	0.821
	ALGM	0.887	0.900	0.887	0.888
	CSFND	0.917	0.917	0.917	<u>0.929</u>
	ChatGLM2-6B	0.892	0.902	0.892	0.892
	LLaMA2-7B	0.908	0.914	0.908	0.908
	BREAK	**0.955**	**0.956**	**0.955**	**0.955**
	Improve(%)	3.355	3.128	3.355	2.799

- **BERT** [14] is a pre-trained language model, and we fine-tune its last two layers for fake news detection. In this paper, BERT is utilized to simulate the performance of fake news detection when only sequential features are available.
- **GAT+2 Attn Heads** [28] represents news sentences as a fully connected graph with randomly initialized node features to detect fake news.
- **SAFE** [40] detects fake news by exploring the similarity between textual and visual information.
- **HMCAN** [22] devises a hierarchical multi-modal attention network to learn a multi-modal news representation.
- **CAFE** [3] aligns cross-modal features and detects fake news by cross-modal ambiguity.
- **GET** [33] treats news and evidence as word-level graphs, respectively. Capturing the long-range semantics to improve fake news detection.
- **MRML** [20] extracts multi-modal relationships and detects multi-modal rumors based on deep metric learning.
- **ALGM** [5] proposes a framework based on the Markov random field and fuses cross-modal features by ambiguity.
- **CSFND** [19] devises an unsupervised fake news detection framework to capture the relationships between news semantic feature space and fake news decision space.

Table 3: The generalization performance of BREAK on fake news datasets with evidence: Snopes and PolitiFact-S. The best performances are in bold, and the runners-up are underlined.

Dataset	Methods	Acc.	Prec.	Rec.	F1
Snopes	BERT	0.766	0.753	0.766	0.741
	GET	0.835	0.838	0.835	<u>0.836</u>
	ChatGLM2-6B	<u>0.843</u>	<u>0.851</u>	<u>0.843</u>	0.829
	LLaMA2-7B	0.836	0.834	0.836	0.827
	BREAK	**0.860**	**0.858**	**0.860**	**0.859**
	Improve(%)	2.017	0.823	2.017	2.751
PolitiFact-S	BERT	0.565	0.563	0.563	0.543
	GET	0.682	0.684	0.682	<u>0.679</u>
	ChatGLM2-6B	0.676	0.675	0.676	0.675
	LLaMA2-7B	**0.710**	**0.711**	**0.710**	**0.708**
	BREAK	<u>0.709</u>	<u>0.708</u>	<u>0.709</u>	**0.708**
	Improve(%)	-0.141	-0.422	-0.141	0.000

- **ChatGLM2-6B** [35] and **LLaMA2-7B** [27] are two large language models (LLMs) with 6B and 7B parameters, respectively. We utilize the same training datasets to **fine-tune** them (using Lora [9]) as the baselines for comparison.

Where GAT+2 Attn Heads, GET, and ALGM leverage the GNN model to improve fake news detection. Meanwhile, LLMs are theoretically suitable for long text modeling since they are trained to handle long sequences of tokens with billions of parameters. In other words, LLMs are strong and competitive baselines, but they come with higher fine-tuning costs and require more advanced fine-tuning skills compared to traditional methods.

4.3 Implementation Details

We partition each dataset into training, validation, and testing sets in an 8:1:1 ratio. We fine-tune the last two layers of $BERT_s$ and $BERT_g$ and the last fully connected layer of $ResNet50$. For the GossipCop, PolitiFact, Snopes, and PolitiFact-S datasets, BREAK is trained in the batch size of 8, 8, 64, and 64 with a hyperparameter β of 0.1, respectively. The learning rates of inner and outer levels are set as 0.1 and 0.00001 separately. Moreover, the dimensions of d and h are set as 768 and 128, respectively. The early-stopping patience is determined to be 8, and Adam [15] is employed as the optimizer. For the metrics, we utilize the weighted accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 score to alleviate the influence of unbalanced datasets.

4.4 Performance Comparison (RQ1)

To evaluate the performance of BREAK on content-only (early) fake news detection, we compare it with seven advanced baselines on the GossipCop and PolitiFact datasets, as shown in Table 2.

BREAK achieves the highest performance across all metrics, leading to a notable improvement of 3.69% and 3.47% in F1 scores compared to sub-optimal results (excluding LLMs) on the GossipCop and PolitiFact datasets, respectively. This improvement underscores the effectiveness of BREAK in fully exploring news content.

Moreover, among all *traditional* baselines, the "GAT+2 Attn Heads" method exhibits competitive results on the GossipCop dataset and the PolitiFact dataset. This finding validates the effectiveness of

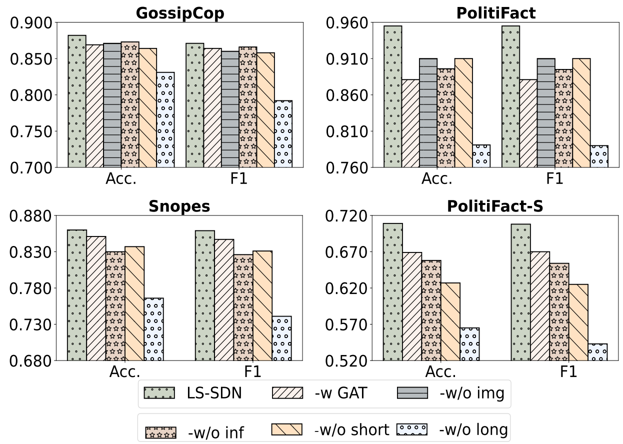


Figure 3: Ablation study on four datasets.

sentence-level graph construction. However, "GAT+2 Attn Heads" falls short when competing with BREAK and novel content-based methods like CSFND, as it initializes node features randomly, overlooking sentence semantics and potential graph noise during the optimization process.

Meanwhile, transformer-based methods like HMCAN, MRML, and CSFND focus on short-range semantics while neglecting some long-range semantics of text. As for CAFE, we observe that it performs significantly worse than other baselines on the PolitiFact dataset. This phenomenon may caused by CAFE detecting fake news through cross-modal ambiguity, but images in PolitiFact are insufficient. Different from CAFE, even though ALGM is also based on cross-modal ambiguity, it further incorporates Markov random fields and semi-supervised settings to improve detection performance.

For LLM baselines, LLaMA2-7B exhibits superior detection results compared to ChatGLM2-6B. We attribute this improvement to the fact that LLaMA2-7B is trained on a larger English corpus than ChatGLM2-6B. Even though LLMs show promising detection performance, they require more fine-tuning costs and are unstable across various datasets. However, our BREAK outperforms them to varying degrees, relying only on a much smaller pre-trained language model, BERT, and a stable hyperparameter across various datasets (details about the hyperparameter can be found in Section 4.7).

4.5 Generalization Exploration (RQ2)

In a real application, **clear** evidence sometimes is available, e.g., the official announcement about the news event. Therefore, we investigate the detection performance of BREAK when such evidence is accessible. Specifically, we conduct experiments on Snopes and PolitiFact-S datasets, in which the evidence directly pertains to the news under consideration. The results are outlined in Table 3.

We compare BREAK with BERT and three strong baselines: the state-of-the-art method on these two datasets (GET) and two LLMs. The results in Table 3 demonstrate that our BREAK exhibits prominent performance, boosting the F1 score by 2.75% compared to the runners-up results on Snopes, and drawing near the detection results

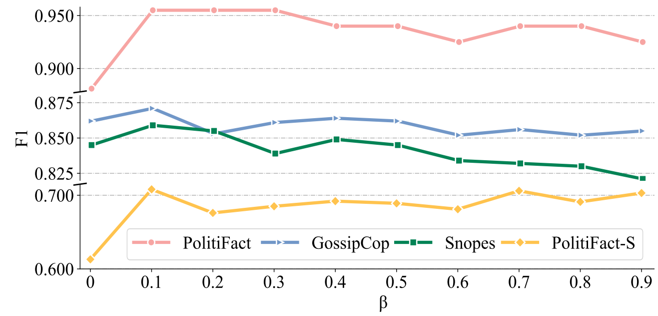


Figure 4: Hyperparameters sensitivity with regard to β.

produced by LLaMA2-7B on PolitiFact-S. Particularly, GET outperforms some baselines to various degrees in F1 score, benefiting from the long-range semantics captured by its word-level graph. However, the word-level graph used in GET overlooks some semantics we discussed in Section 1, while BREAK models and denoises all semantics appropriately.

Therefore, the results on four datasets demonstrate BREAK's generality and stability.

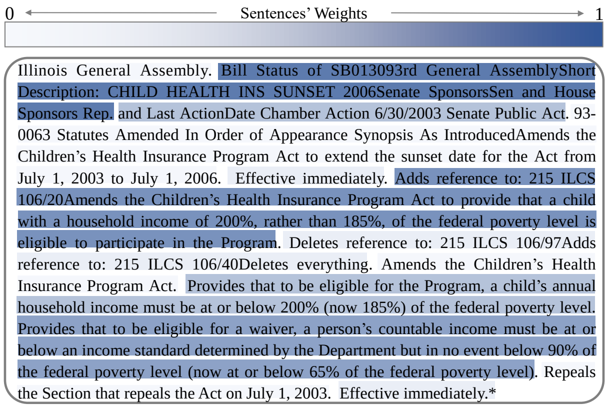
4.6 Ablation Study (RQ3)

To assess the necessity of each part in BREAK, we compare BREAK with its four variants: "-w GAT," "-w/o img," "-w/o inf," "-w/o seq," and "-w/o gra." Specifically, "-w GAT" represents the variant that utilizes graph attention network (GAT) as the denoise module and graph encoder, "-w/o img" excludes visual content in the GossipCop and PolitiFact datasets, "-w/o inf" omits the sequential semantic integration mechanism and edge weight inference mechanism, "-w/o seq" disregards the sequence encoder and only employing GCN for fake news detection (i.e., sequential semantics), and "-w/o gra" means without the graph structure and only utilizing BERT as the detection model.

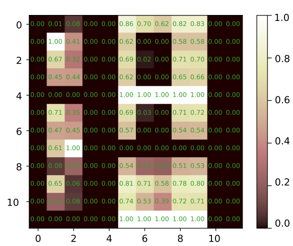
The comparative results are illustrated in Figure 3, highlighting that the absence of any part of BREAK results in sub-optimal performance. In detail, "-w GAT" indicates that the attention mechanism cannot handle such complex semantic dependencies well. Moreover, the performances of "-w/o inf" on all datasets are very close those of "-w/o seq," as "-w/o inf" considers noise semantics during semantic integration. These experimental phenomena demonstrate the effectiveness of the inner denoising process. Meanwhile, both "-w/o seq" and "-w/o gra" show sub-optimal results, with "-w/o gra" performing the worst, indicating that integrating broad-range semantics from the graph and sequence is necessary, and that structural semantics are more crucial for comprehensive news representation modeling.

4.7 Sensitivity of Hyperparameter β (RQ4)

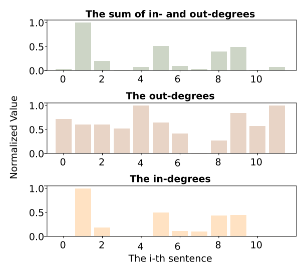
In the BREAK model, the hyperparameter β balances the combination of structural and sequential semantics. We analyze its impact on fake news detection performance, with results shown in Figure 4. BREAK ***consistently*** achieves optimal results across four datasets with $\beta = 0.1$, demonstrating its necessity and generalizability. Meanwhile, significant improvements are observed in both PolitiFact and PolitiFact-S datasets, consistently surpassing results with β values



(a) The visualization of sentences' weights.



(b) Denoised edge weights.



(c) Degrees of nodes.

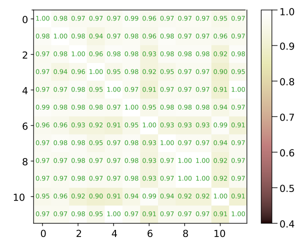
Figure 5: A case study of structure denoising. (a) represents the visualization of sentences' weights. (b) represents the weight of each edge between sentences. (c) depicts the normalized in- and out-degrees of each node.

from 0.1 to 0.9 compared to $\beta = 0$. However, larger values of β lead to various degrees of performance degradations on all datasets, which incorporate *more noise* from the fully connected graph to short-range semantics and thereby hinder the denoise process.

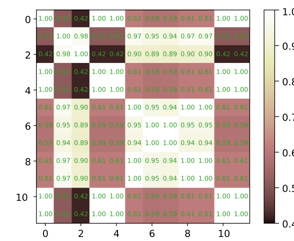
4.8 Case Study on Structure Denoising and Feature Diversification (RQ5)

We evaluate the effectiveness of BREAK in structure denoising and feature diversification through a case study. Specifically, we visualize a news article to demonstrate its edge weights and the similarity between sentences (node features), as depicted in Figure 5 and Figure 6.

Structure Denoising. The edge weights learned by BREAK in two directions are illustrated in Figure 5(b), where rows (columns) indicate the forward (backward) direction. In BREAK, the noise semantics are weakened to close to 0, and the key semantics are enhanced to around 1. Moreover, these edge weights also exhibit certain interpretability for the detection results. Specifically, we sum all the out-degrees and in-degrees of a node (sentence) as its weight, and the normalized results are shown in Figure 5(c). Surprisingly, we observe that node weights are primarily determined by in-degrees (backward connections). Additionally, a majority of nodes disseminate their information to other nodes, exhibiting high out-degrees.



(a) Original similarity



(b) Diversified similarity

Figure 6: A case study of node features diversification. (a) indicates the original similarity of news sentences. (b) denotes the similarity of news sentences learned by BREAK.

From this observation, we can conclude that sentences with high in-degrees are likely to be topic sentences, as other sentences revolve around these focal points. Meanwhile, as depicted in Figure 5(a), BREAK assigns distinct weights to sentences, revealing key facts crucial for fake news detection. In detail, sentences with higher weights convey the facts that the proposal of a children's health bill and the resulting modifications in eligibility criteria. Moreover, we observe that sentences with higher weights predominantly function as factual statements, resulting in higher in-degrees. Conversely, sentences with lower weights play a supporting role, leading to higher out-degrees. It's noteworthy that news articles tend to reiterate facts, such as "effective immediately" and "change from 185% to 200%", contributing to the high similarity among sentences.

Feature Diversification. The cosine similarity between news sentences pre- and post-training is presented in Figure 6(a) and Figure 6(b). The initial similarity obtained by the BERT model is notably high, with values consistently exceeding 0.9. In contrast, the sentence features learned by our BREAK are more distinct. This can effectively mitigate the over-smoothing problem during graph aggregation and aid in identifying key news sentences.

5 CONCLUSION

In this paper, we propose BREAK, a network devised to extract and integrate the broad-range semantics while avoiding noise incorporation. BREAK models broad-range semantics as a fully connected graph and implements dual denoising modules under a bi-level optimization paradigm to mitigate both structural and feature noise. These two modules effectively eliminate irrelevant semantic interrelations and diversify semantic features, respectively. Specifically, at the inner level, we introduce a sequence-based structure to obtain the sequential semantics and the lower bound of structure. Moreover, a structural and sequential semantic integration mechanism and an edge weight inference mechanism are devised to achieve structural denoising by narrowing the structure bounds. At the outer level, we employ KL-divergence to align the graph and sequence encoders, thereby diversifying semantic features and integrating broad-range semantics for fake news detection. Extensive experiments on both content-only and clear-evidence-accessible scenarios demonstrate the superiority of BREAK in fake news detection.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–236.

[2] Margarita Bugueño and Gerard de Melo. 2023. Connecting the Dots: What Graph-Based Text Representations Work Best for Text Classification using Graph Neural Networks?. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8943–8960. https://doi.org/10.18653/v1/2023.findings-emnlp.600

[3] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-Modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2897–2905. https://doi.org/10.1145/3485447.3511968

[4] Yasan Ding, Bin Guo, Yan Liu, Yao Jing, Maolong Yin, Nuo Li, Hao Wang, and Zhiwen Yu. 2025. EvolveDetector: Towards an evolving fake news detector for emerging events with continual knowledge accumulation and transfer. *Information Processing & Management* 62, 1 (2025), 103878.

[5] Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, Xiaowen Su, and Di Jin. 2023. A generalized deep markov random fields framework for fake news detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 4758–4765.

[6] Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, and Jingrui He. 2022. DISCO: comprehensive and explainable disinformation detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4848–4852.

[7] Lukas Galke and Ansgar Scherp. 2022. Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4038–4051. https://doi.org/10.18653/v1/2022.acl-long.279

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

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

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

[12] Yinqiu Huang, Min Gao, Jia Wang, Junwei Yin, Kai Shu, Qilin Fan, and Junhao Wen. 2023. Meta-prompt based learning for low-resource false information detection. *Information Processing & Management* 60, 3 (2023), 103279.

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

[14] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*. PMLR, 15524–15543.

[17] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[18] Giannis Nikolentzos, Antoine Tixier, and Michalis Vazirgiannis. 2020. Message passing attention networks for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8544–8551.

[19] Liwen Peng, Songlei Jian, Zhigang Kan, Linbo Qiao, and Dongsheng Li. 2024. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Information Processing & Management* 61, 1 (2024), 103564.

[20] Liwen Peng, Songlei Jian, Dongsheng Li, and Siqi Shen. 2023. MRML: multimodal rumor detection by deep metric learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[21] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1003–1012.

[22] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).

[23] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.

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

[25] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[26] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*. 2789–2797.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[28] Vaibhav Vaibhav, Raghuram Mandyam Annasamy, and Eduard H. Hovy. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)* (2019). https://api.semanticscholar.org/CorpusID:204904180

[29] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 18–22.

[30] Jia Wang, Min Gao, Yinqiu Huang, Kai Shu, and Hualing Yi. 2023. FinD: Fine-grained discrepancy-based fake news detection enhanced by event abstract generation. *Computer Speech & Language* 78 (2023), 101461.

[31] Jiaying Wu and Bryan Hooi. 2023. DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2582–2593.

[32] Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. MSynFD: Multi-hop Syntax aware Fake News Detection. In *Proceedings of the ACM on Web Conference 2024*. 4128–4137.

[33] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*. 2501–2510.

[34] Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 791–800.

[35] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=-Aw0rrrPUF

[36] Hao Zhang, Zonglin Li, Sanya Liu, Tao Huang, Zhouwei Ni, Jian Zhang, and Zhihan Lv. 2023. Do Sentence-Level Sentiment Interactions Matter? Sentiment Mixed Heterogeneous Network for Fake News Detection. *IEEE Transactions on Computational Social Systems* (2023).

[37] Litian Zhang, Xiaoming Zhang, Chaozhuo Li, Ziyi Zhou, Jiacheng Liu, Feiran Huang, and Xi Zhang. [n. d.]. Mitigating Social Hazards: Early Detection of Fake News via Diffusion-Guided Propagation Path Generation. In *ACM Multimedia 2024*.

[38] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*. 3465–3476.

[39] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 334–339. https://doi.org/10.18653/v1/2020.acl-main.31

[40] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Proceedings* abs/2003.04981 (2020),

354–367.
[41] Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Juergen Kurths. 2024. Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4652–4663.

A APPENDIX

Some prefix knowledge involved in Eq. (12) as follows:

Self Information. Given a random variable $a \in \mathcal{A}$. Its self information can be written as $S(a) = -log\ \mathbb{P}(a)$, where $\mathbb{P}$ indicates the distribution of a.

Entropy. The entropy of a is defined as the expectation of $S(a)$: $H(a) = \mathbb{E}_a[S(a)] = -\sum_{a\in\mathcal{A}} \mathbb{P}(a) log\ \mathbb{P}(a)$.

KL Divergence. KL divergence measures the discrepancies between two distributions. Specifically, given random variable a and its two distributions (true and predicted) $\mathbb{P}(a)$ and $\mathbb{Q}(a)$, the KL divergence can be written as $KL(\mathbb{P}(a), \mathbb{Q}(a)) = \sum_{a\in\mathcal{A}} \mathbb{P}(a) \frac{\mathbb{P}(a)}{\mathbb{Q}(a)}$, which can be utilized to measure how the predicted distribution $\mathbb{Q}(a)$ close to the true distribution $\mathbb{P}(a)$.

Mutual Information. The mutual information $I(a; b)$ quantifies the reduction in uncertainty about variable a when the value of another variable b is known, i.e., assesses the degree of dependence or correlation between a and b. Formally, the mutual information can be described as $I(a; b) = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \mathbb{P}(a, b) \log \frac{\mathbb{P}(a,b)}{\mathbb{P}(a)\mathbb{P}(b)}$.

The inner structure denoising process aims to extract key sentences or images for effective fake news detection. Utilizing known news labels in the training set, we seek to maximize the mutual information between news representations and labels. This maximization assists the denoising process in excavating the most label-related information and eliminating non-relevant noise.

By the definition of mutual information, we can formally write this optimization objective as Eq.(10). Following the previous work [16], we give the detailed derivation of transforming Eq. (10) into Eq. (12). BREAK detects fake news by the representation of news E, i.e., $\mathbb{P}(Y|E) = \frac{\mathbb{P}(E,Y)}{\mathbb{P}(E)}$. However, $\mathbb{P}(Y|E)$ is intractable since no ground-truth label to model the relationship between E and Y. Therefore, we introduce a variational approximation $\mathbb{P}_{\phi]}(Y|E)$ for $\mathbb{P}(Y|E)$. Accordingly, we acquire the lower bound of Eq. (10) as follows:

$$I(E; Y) = \sum_{E,Y} \mathbb{P}(E, Y) log\ \frac{\mathbb{P}(Y|E)}{\mathbb{P}(Y)} \tag{15}$$

$$= \mathbb{E}_{E,Y} \left[log\ \frac{\mathbb{P}(Y|E)}{\mathbb{P}(Y)} \right] \tag{16}$$

$$= \mathbb{E}_{E,Y} \left[log\ \frac{\mathbb{P}_\phi(Y|E)}{\mathbb{P}(Y)} \right] + \mathbb{E}_{E,Y} \left[KL(\mathbb{P}(Y|E), \mathbb{P}_\phi(Y|E)) \right] \tag{17}$$

$$\geq \mathbb{E}_{E,Y} \left[log\ \frac{\mathbb{P}_\phi(Y|E)}{\mathbb{P}(Y)} \right] \tag{18}$$

$$= \mathbb{E}_{E,Y} \left[log\ \mathbb{P}_\phi(Y|E) \right] - \mathbb{E}_{E,Y} \left[log\ \mathbb{P}(Y) \right] \tag{19}$$

$$= \mathbb{E}_{E,Y} \left[log\ \mathbb{P}_\phi(Y|E) \right] + H(Y), \tag{20}$$

where the KL divergence in step (17) is utilized to measure the difference between the true distribution $\mathbb{P}(Y|E)$ and the variational approximation $\mathbb{P}_\phi(Y|E)$. Ultimately, we acquire Eq. (12) by inverse the lower bound as $-I(E; Y) \leq -\mathbb{E}_{E,Y} \left[log\ \mathbb{P}_\phi(Y|E) \right] - H(Y)$.