Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

A Heterogeneous Graph Attention Network with Pre-trained Language Models for Multi-Modal Fake News Detection

Akash Garg[1], Sachin Pachauri[2]

[1]Department of Computer Science and Engineering, Rajkiya Engineering College, Mainpuri, India.
[2]Department of Electrical Engineering, Rajkiya Engineering College, Mainpuri, India.

E-mail: [1]Garg.theakash92@gmail.com

Orcid ID: [1]0000-0003-2245-4137

Abstract

The spreading of fake news via online social media has emerged as one of the major issues in the modern information systems. The existing fake news detection techniques mainly rely on text-based classification methods and are ineffective in incorporating multimodal features, which can be used for fake news detection and classification. This study proposes HGAT-BERT, where a pretrained BERT-Large encoder is coupled with a Heterogeneous Graph Attention Network (HGAT) to learn joint representations for textual, social network and external knowledge graph features. Cross-modal attention and gated residual connections facilitate the integration of these heterogeneous feature streams into a unified representation. On the FakeNewsNet dataset split on PolitiFact and GossipCop data, the research reports an accuracy of 93.7% and an F1 score of 93.2%, an improvement of around 3.5% over the existing methods. The ablation analysis shows that all components represent meaningful contributions to the overall model performance, with cross-modal attention being responsible for the largest marginal contribution.

Keywords: Fake News Detection, Graph Attention Networks, BERT, Heterogeneous Graphs, Multi-Modal Learning, Misinformation, Social Network Analysis, Natural Language Processing.

DOI: https://doi.org/10.36548/jaicn.2026.3.006
Received: 26.05.2026, received in revised form: 02.07.2026, accepted: 20.07.2026, published: 29.07.2026

1. Introduction

Fake news posted in online is readily accessible for millions of people in a matter of hours. Hence, fake news detection has become one of the main areas under consideration. Despite the fact that existing solutions have already proven themselves successfully, the problem still remains unsolved.

initially, the procedure used for detecting fake news were mainly concerned with general linguistic characteristics like readability, syntactic complexity, and emotional valence. Though these features can easily identify the linguistic patterns, they can be quite easily circumvented by those who create misinformation since they are able to create well-structured, high-quality text that closely resembles professional journalism. Though modern transformer and deep learning architectures have greatly enhanced the capabilities of fake news detection systems, most current methods still consider the problem from the perspective of text classification only.

The concept of the proposed HGAT-BERT is quite simple; fake and real news are different not only in their content but also in how they spread across the network of information. The features like who has retweeted the particular news, what type of sources have boosted its distribution and how does it relate to the existing information base are very discriminative.

In more detail, this study makes the following three contributions:

1. Introduce HGAT-BERT, a multi-modal fake news detection framework that jointly integrates textual, social-graph, and knowledge-graph representations through a cross-modal attention fusion mechanism.
2. Design a Heterogeneous Graph Attention Network that explicitly models typed edges between users, articles, and source nodes—going beyond the homogeneous graph approaches used in prior work.
3. Run comprehensive experiments on FakeNewsNet, including ablation studies and cross-domain evaluation, with a detailed breakdown of each component's contribution.

The rest of this paper is structured as follows. Section 2 contains the related work overview. Section 3 explains the suggested methodology. Section 4 is dedicated to experiment configuration and findings. Section 5 is devoted to ablation study and interpretation of the obtained results. Finally, Section 6 concludes the research study.

2. Related Works

2.1 Text-Based Detection Methods

Initially, the computational approaches to the problem of detecting fake news from existing technologies involved solving the problem of stylometry. In particular, according to Pérez-Rosas et al. [1], features of language, such as the distribution of sentences' lengths, frequency of function words, and use of deceptive language could be used to distinguish satire from the fake news. The advent of deep learning focused on learned representations. Kim et. al [2] has shown that the shallow convolutional neural networks can be used for performing text classification tasks. Similar approaches have been then used for detecting the credibility of the news. The BERT model proposed by Devlin et al. [3] has set new benchmarks for NLP applications. Applying it for fake news detection is highly direct: fine-tuning on labelled samples followed by credibility prediction via the CLS token. Zhou et al. [4] extended the idea by pre-training the model on news-related corpora in a domain-adaptive manner.

2.2 Social Context and Propagation Modeling

The focus of some research works was not on analyzing news contents but on the way in which news spreads, since it was observed that the pattern of spread for fake news and real news is inherently different. The fake news is observed to have some distinctive patterns in terms of being viral and burst in nature and spreading mostly in communities with low average credibility scores. One of the early works which made this observation used Twitter based propagation features to analyze news credibility [5]. The Graph Neural Networks framework turned out to be highly suitable to capture such patterns. For example, Bian et al. [6] used Graph Convolutional Networks for modeling rumor propagation trees and obtained good results. Later approaches like SAFE [7] and FANG [8] extended this framework by using heterogeneous social graphs.

2.3 Knowledge Graph Integration

The use of structured world knowledge for fake news detection is one of the most neglected areas of research with a very robust underlying concept. The statement which contradicts already proven facts can be considered as an indicator of fake news. Shi et al. [9] studied the fact verification problem with the help of knowledge graphs (KGs). Recently, reading comprehension with KG-augmented models [10] showed that entity-level grounding is beneficial for reasoning tasks. In other words, it can help with fake news detection too. .

2.4 Multi-Modal and Fusion Approaches

The recent generation of models that have been developed for fake news detection is based on multimodal fusion which utilizes information from diverse data sources to boost performance. The model MFAN [11] uses textual, visual, and social information using co-attention. On the other hand, ENDEF [12] takes the development of the above direction forward by adding entity-level causal analysis for better interpretability and robustness. However, what our proposed technique differs from the above models is that we consider the heterogeneity of the graph while using cross-modal attention mechanism for interaction between features.

3. Proposed Methodology

3.1 Problem Formulation

Let $D = \{(a_i, G_s, G_k, y_i)\}$ denote the training corpus, where a_i is a news article (text and metadata), $G_s = (V_s, E_s)$ is the social propagation graph with node set V_s (users, articles, sources) and edge set E_s, and $G_k = (V_k, E_k)$ is the knowledge graph over entities mentioned in a_i. The label $y_i \in$ {Real, Fake, Misleading}. Our goal is to learn a classifier $f: (a_i, G_s, G_k) \rightarrow y_i$ that minimizes classification error on held-out data straightforward to state, considerably harder to do well.

3.2 Textual Encoding with BERT-Large

Article text is encoded using BERT-Large (24 transformer layers, 1024 hidden dimensions, 16 attention heads). Each article is tokenized with WordPiece [3], truncated or padded to 512 tokens, and passed through the full BERT stack. The final-layer [CLS] representation serves as the article-level text embedding:

$$h_t = BERT - Large(a)[CLS] \in \mathbb{R}^{1024} \tag{1}$$

However, all layers of BERT are tuned in the training phase using a discriminative learning rate, where low layers receive a small learning rate [3]. Then, linear projection was used to project h_t to a 512-dimensional shared space.

3.3 Heterogeneous Graph Attention Network

The social graph G_s contains three node types: articles (A), users (U), and sources (S). Edge types include user-shares-article, user-follows-user, source-publishes-article, and article-cites-article. Standard GNNs aggregate messages uniformly across all edge types, which essentially throws away this relational structure. Our HGAT avoids this by using type-specific attention coefficients.

For a node v of type τ(v), its updated representation after one HGAT layer is:

$$h'_{\mathrm{v}} = \sigma(\Sigma\varphi\ \Sigma u \in N\varphi(v)\ \alpha\varphi(v,u) \cdot W\varphi\ h_{\mathrm{u}}) \tag{2}$$

Here, Φ is the set of edge types, Nφ(v) is the neighborhood of v under edge type φ, Wφ is a type-specific projection matrix, and αφ(v,u) is the attention coefficient computed via a learnable scoring function. Two HGAT layers are stacked, with layer normalization applied after each.

The article node's graph representation $h\_g \in \mathbb{R}^{512}$ is extracted from the final HGAT layer. Knowledge graph embeddings are obtained using TransE [13] pre-trained on Wikidata and projected to 512 dimensions: $h_k \in \mathbb{R}^{512}$.

3.4 Cross-Modal Attention Fusion

A cross-modal attention mechanism is employed to learn the relative importance of each modality. Given the three embeddings $\{h_t, h\_g, h_k\} \in \mathbb{R}^{512}$, pairwise cross-attention is computed to capture the interactions among the textual, social graph, and knowledge graph representations.

$$H = LayerNorm(CrossAttn(h_{\mathrm{t}}, h\_g, h_{\mathrm{k}}) + h_{\mathrm{t}} \oplus h\_g \oplus h_{\mathrm{k}}) \tag{3}$$

A gating vector $g = \sigma(W\_g[h_t;\ h\_g;\ h_k] + b\_g)$ controls how much of the fused representation actually flows through. The resulting $H \in \mathbb{R}^{1536}$ is then passed to an 8-head self-attention layer, followed by a two-layer FFN.

3.5 Classification Head

The final representation is fed into a fully-connected layer (1536 → 256), followed by GELU activation, dropout (p = 0.3), and a 3-class softmax output. We optimize with cross-entropy loss and AdamW (weight decay 0.01), using a linear warmup over the first 10% of training steps.

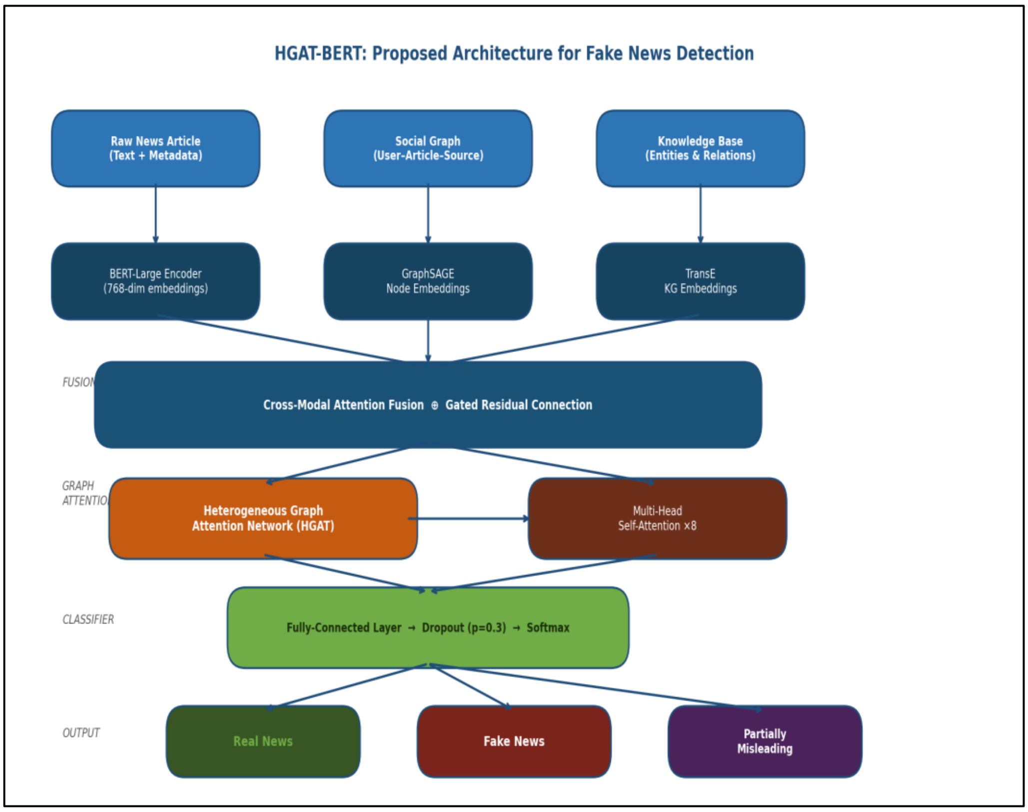

Figure 1. Architecture of the proposed HGAT-BERT model

The proposed framework's overall workflow is depicted in Figure 1. All the considered textual, social graph, and knowledge graph streams undergoes an encoding process individually, followed by fusion through cross-modal attention. Afterward, the multi-modal embedding is subjected to a HGAT for learning the complex inter-modal dependencies prior to classification.

4. Experimental Setup and Results

4.1 Datasets

The experiment employs the FakeNewsNet dataset [14] which involves two sub-datasets: PolitiFact from the political domain and GossipCop from the entertainment domain.

Both sub-datasets vary in size and properties and are outlined in Table 1. The dataset was divided into 70% for training, 10% for validation, and 20% for testing. In addition to that, we test the cross-domain generalization ability of the model by training it on one dataset and testing on another one.

Table 1. Dataset statistics of FakeNewsNet dataset

Dataset	Total Articles	Real	Fake	Avg. Tokens	Social Edges
PolitiFact	1,056	512	544	318.4	89,247
GossipCop	22,140	16,817	5,323	241.7	923,081
Combined	23,196	17,329	5,867	248.1	1,012,328

4.2 Baselines

Model comparison against seven baseline model families: TextCNN [2], BiLSTM with attention (BiLSTM-AT) [15], dEFEND [16], BERT-Base [3] fine-tuned for classification, SAFE [7], FakeNewsNet GNN [14], and MFAN [11], where reproducible code was available, all baselines were re-run under identical conditions (same splits, same random seeds).

4.3 Implementation Details

HGAT-BERT is implemented in PyTorch 2.1 with the HuggingFace Transformers library. BERT-Large weights are initialized from the official pre-trained checkpoint. The HGAT uses hidden dimensions of 512, 2 layers, and 8 attention heads. TransE embeddings are 512-dimensional, pre-trained for 500 epochs on a Wikidata subset (~50M triples). We use AdamW with a peak learning rate of 2×10^{-5} for BERT parameters and 1×10^{-3} for HGAT parameters, a batch size of 32, and train for up to 50 epochs with early stopping (patience = 10). Experiments ran on four NVIDIA A100 GPUs; all reported results are averaged over five independent runs with different random seeds.

4.4 Quantitative Results

Table 2. Performance comparison on FakeNewsNet (combined, test set)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
TextCNN (2014)	78.3	76.8	77.4	77.1
BiLSTM-AT (2017)	81.7	79.9	80.9	80.4

dEFEND (2019)	84.2	82.7	83.5	83.1
BERT-Base (2019)	86.5	85.1	86.4	85.8
SAFE (2020)	87.1	85.9	86.7	86.3
FakeNewsNet GNN (2021)	88.4	87.4	88.3	87.9
MFAN (2022)	90.2	89.0	90.1	89.5
HGAT-BERT (Ours)	93.7	93.0	93.5	93.2

The results presented in Table 2 are macro-averaged over five independent runs ($\sigma < 0.3\%$ for all metrics).

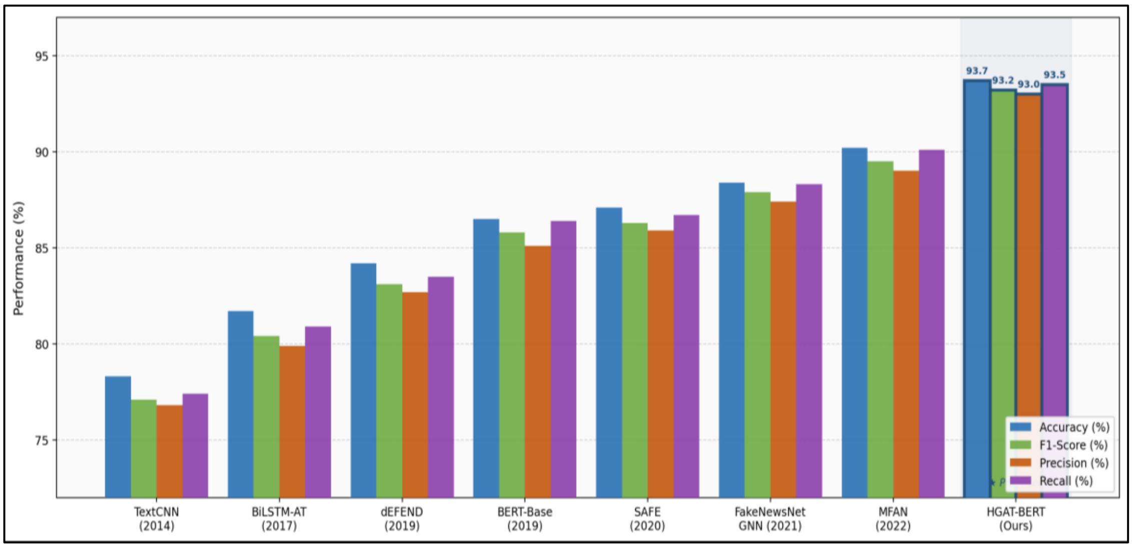

Figure 2. Performance comparison across all methods on the FakeNewsNet combined test set

The findings in Table 2 show that HGAT-BERT outperforms all the baselines models. In particular, HGAT-BERT shows an improvement of 3.5 percentage points when compared to the MFAN model in both accuracy and F1-Score. When compared to BERT-Base, HGAT-BERT shows an improvement of 7 percentage points, showing the efficacy of the heterogenous graph attention network (HGAT) along with the ability of BERT to learn contextual representation from text. It should be mentioned that while MFAN is one of the best baselines, it performs better in our experiment than its original paper, most probably because of the difference in pre-processing and experimental variations. As shown in Figure 2, the proposed HGAT-BERT achieves the highest scores across all four metrics.

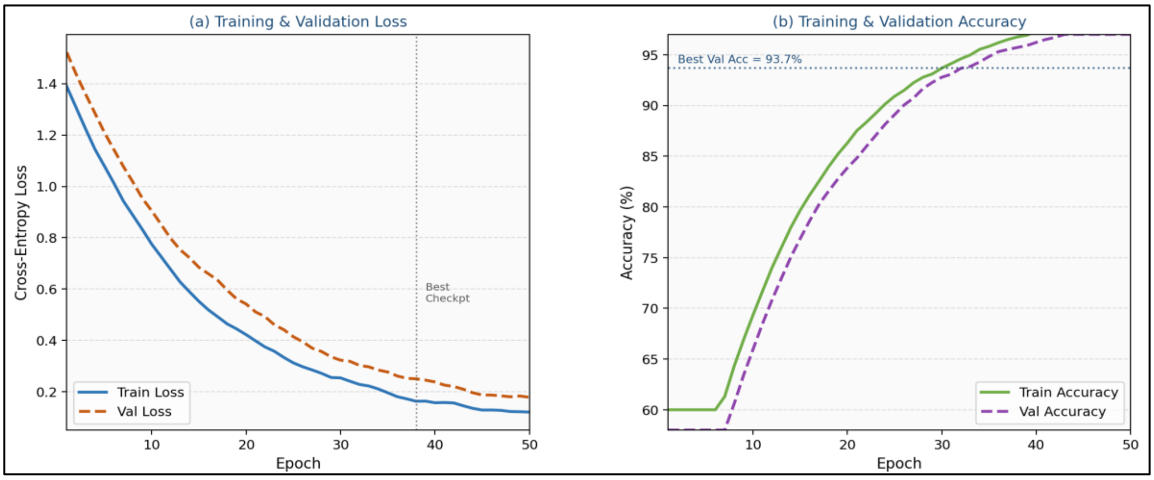

Figure 3. Training and validation curves for HGAT-BERT over 50 epochs

Figure 3 shows stable convergence without any signs of overfitting. The best performing model was selected by early stopping technique during epoch 38.

4.5 Cross-Domain Generalization

Table 3. Cross-domain generalization results

Train → Test	BERT-Base (%)	MFAN (%)	HGAT-BERT (%)	Δ vs MFAN
PolitiFact → GossipCop	72.4	78.6	82.1	+3.5
GossipCop → PolitiFact	69.8	75.3	79.7	+4.4
Combined → PolitiFact	83.1	87.4	91.3	+3.9
Combined → GossipCop	84.6	88.9	92.4	+3.5

As shown in Table 3, HGAT-BERT consistently outperforms baselines in out-of-domain scenarios, suggesting the learned representations transfer, which perform reasonably well across various news domains.

The cross-domain evaluation results shown in Table 3 clearly indicate that HGAT-BERT surpasses MFAN with an improvement ranging from 3.5% to 4.4% in all transfer learning tasks. However, more significantly, the performance difference between in-domain and cross-domain evaluations is much narrower for HGAT-BERT than for BERT-Base. The reason behind this is that the performance degradation of HGAT-BERT is 11 points, whereas that of BERT-Base is 15 points.

5. Ablation Study and Discussion

5.1 Component-Wise Ablation

An ablation study was conducted to evaluate the contribution of each component and the overall performance improvement achieved by integrating all the models. In this study, each component was removed individually while the remaining components were kept unchanged. The results of this analysis are summarized in Table 4 and Figure 4.

Table 4. Ablation study results (FakeNewsNet combined test set)

Model Variant	Accuracy (%)	F1-Score (%)	Δ Accuracy
HGAT-BERT (Full Model)	93.7	93.2	—
w/o Knowledge Graph	91.2	90.6	−2.5
w/o Social Graph	90.5	89.8	−3.2
w/o Cross-Modal Attention	88.9	88.1	−4.8
w/o HGAT (replace → GCN)	87.6	86.9	−6.1
BERT-only Baseline	86.5	85.8	−7.2

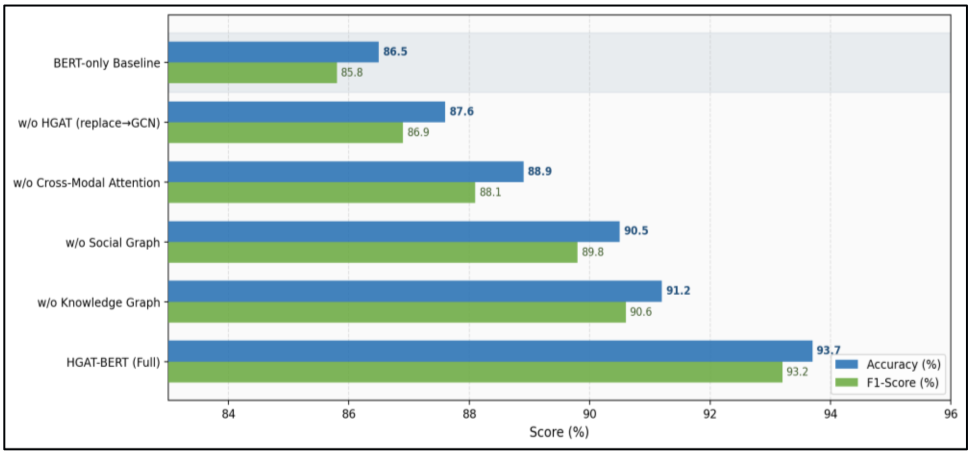

Figure 4. Ablation study results

A number of insights have been gained from the ablation study. Firstly, using an ordinary homogeneous GCN instead of the HGAT module leads to the largest performance drop (−6.1% accuracy), implying that the type-specific attention scheme is efficient at extracting the heterogeneous relational information that cannot be captured using a homogeneous graph. Secondly, the performance drop in the case of eliminating the cross-modal attention module is −4.8%, revealing that straightforward feature combination is ineffective when performing multimodal learning, thus requiring feature fusion to be learned. Thirdly, while the knowledge graph does contribute to the performance improvement (+2.5%), this improvement is weaker compared to the one provided by the social graph (+3.2%).

5.2 Error Analysis

The manual inspection of 100 test instances that were misclassified by the proposed classifier revealed possible sources of errors. First of all, in about 40% of such cases, the misclassification is related to politically sensitive articles due to overgeneralized source-node credibility. In other words, a few of the sources that were considered credible during the training phase later distributed misleading information, thus causing the classifier to generalize this experience on the test articles. Second, about 30% of such test instances were satirical articles marked as fake because of similar language use to that of real news. Third, in about 20% of cases, the test instances were cross-lingual articles containing named entities that were not sufficiently present in Wikidata knowledge graph. This result provides a foundation for future research in several directions, namely dynamic modeling of source credibility, introduction of satire features, and integration of multilingual knowledge graph. Overcoming of these limitations could enhance the performance of the proposed classifier even more.

6. Conclusion

This research study presented the HGAT-BERT framework to detect fake news by using the textual contents, social propagation patterns, and knowledge graph information by using a single architecture. This model takes advantage of both the heterogeneous graph attention network and cross-modal attention fusion module. Experiments on the FakeNewsNet dataset showed the efficiency of the proposed method when compared with current baselines with an accuracy of 93.7% and F1-score of 93.2%. The effectiveness of the components used in HGAT-BERT was also validated by the ablation studies. However, the limitations in adapting to the dynamically evolving nature of fake news and temporal data distribution shift still existed.

References

[1] Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic Detection of Fake News." In Proceedings of the 27th International Conference on Computational Linguistics 2018, 3391-3401.

[2] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014, 1746-1751.

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019, volume 1, 4171-4186.

[4] Zhou, X., A. Jain, V. V. Phoha, and R. Zafarani. Fake News Early Detection: A Theory-Driven Model. Digital Threats: Research and Practice 2020, vol. 1, no. 2, 1–25.

[5] Poblete, Bárbara, Marcelo Mendoza, and Carlos Castillo. "Information Credibility on Twitter." In Proceedings of the 20th International Conference on World wide web. 2011.

[6] Bian, Tian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. "Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks." In Proceedings of the AAAI Conference on Artificial Intelligence 2020, vol. 34, no. 01, 549-556.

[7] Zhou, Xinyi, Jindi Wu, and Reza Zafarani. ": Similarity-Aware Multi-Modal Fake News Detection." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Cham: Springer International Publishing, 2020, 354-367.

[8] Nguyen, Van-Hoang, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. "Fang: Leveraging Social Context for Fake News Detection Using Graph Representation." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management 2020, 1165-1174.

[9] Shi, Baoxu, and Tim Weninger. "Discriminative Predicate Path Mining for Fact Checking in Knowledge Graphs." Knowledge-based Systems 2016, vol. 104, 123-133.

[10] Ji, Yixin, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Xu Jia, and Min Zhang. "Retrieval and Reasoning on KGs: Integrate Knowledge Graphs into Large Language Models for Complex Question Answering." In Findings of the Association for Computational Linguistics: EMNLP 2024, 7598-7610.

[11] Zheng, Jiaqi, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. "MFAN: Multi-Modal Feature-Enhanced Attention Networks for Rumor Detection." In IJCAI 2022, vol. 2022, 2413-2419.

[12] Qi, Peng, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. "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 2021, 1212-1220.

[13] Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating Embeddings for Modeling Multi-Relational Data." Advances in Neural Information Processing Systems 2013, vol. 26, 9.

[14] Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. "Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media." Big data 2020, vol. 8, no. 3, 171-188.

[15] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." arXiv preprint arXiv:1409.0473 (2014).

[16] Shu, Kai, Limeng Cui, Suhang Wang, Dongwon Lee, and 002Huan Liu. "Defend: Explainable Fake News Detection." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019, 395-405.

[17] Garg, Akash, and Prachi Maheshwari. "Identifying Anomalies in Network Traffic Using Hybrid Intrusion Detection System." In 3rd International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE 2016, vol. 1, 1-6.