Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA36324

Multimodal Fusion Model for Fake News Detection Combining BERT and Graph Attention Networks

Jiale Wang

Kang Chiao International School, Kunshan, China
15050264080@163.com

Abstract. Fake-news detection needs both semantic and social evidence: text-only systems miss coordinated diffusion, while graph-only systems may confuse virality with falsehood. We propose a framework that combines BERT text embeddings with Graph Attention Network (GAT) representations of propagation graphs. Projected features are aligned by a cross-modal consistency loss and combined through an adaptive attention gate; when graphs are unavailable, the structural branch is masked. FakeNewsNet and Twitter15/16 evaluate full fusion, while LIAR tests text-only operation. On FakeNewsNet, the model reaches 93.4% accuracy, 92.8% Macro-F1, and 0.97 AUC. Ablations support both fusion components, and token- and node-level attribution assists human review.

Keywords: Fake news detection, BERT, Graph attention network, Multimodal fusion, Social propagation

1. Introduction

Social media accelerates access to news but also spreads fabricated or misleading content, with consequences for political discourse, public health, and institutional trust. Fake news includes fabricated stories, manipulated media, propaganda, and other deceptive narratives lacking factual integrity [1]. Manual verification cannot match platform scale, making reliable automated support necessary.

Detection is difficult because misinformation has both semantic and social properties. Deceptive language may use emotional framing or unsupported certainty, while coordinated accounts can create abnormal reposting patterns. Text-only models miss this diffusion evidence; graph-only models can mistake legitimate breaking-news virality for falsehood.

This paper aligns BERT semantic features [2] with GAT propagation features [3], then learns an item-specific fusion weight and a cross-modal consistency objective. The goals are improved accuracy, robust operation when one modality is weak or missing, and auditable predictions based on influential words and propagation nodes.

DOI: 10.54254/2755-2721/2026.BA36324

2. Literature review

2.1. Text-based fake news detection

Classical detectors use lexical, sentiment, readability, and source features. BERT instead learns bidirectional contextual representations and can be fine-tuned for classification [2]; FakeBERT adapts this approach to fake news [7]. However, text-only models may learn topic-specific shortcuts, struggle with professionally written deception, and cannot observe who shared a claim or how it spread.

2.2. Network-based and propagation-oriented approaches

Propagation methods model source posts, replies, retweets, or users as trees or graphs. Structural kernels, Bi-GCN, and GAT capture cascade direction, branching, and influential participants [3-5]. Their limitations are sparse early graphs, restricted interaction data, computational cost, and weak cross-platform transfer. Because rapid diffusion also occurs for true news, structure should complement rather than replace content.

2.3. Multimodal and cross-modal fusion techniques

Multimodal research combines text with images, profiles, knowledge, or social context. For text-social fusion, GCAN highlights source words and suspicious retweeters [6], while FANG learns representations from social context [8]. The central problem is not simple concatenation: channels differ in scale, noise, and availability. Our framework therefore projects them into a shared space, controls modality dominance with attention, and masks missing graph input.

3. Methodology

The model has separate text and graph encoders, a shared projection space, and an adaptive classifier. Each input is encoded independently; aligned features are fused before label prediction.

3.1. Textual feature extraction

BERT-base-uncased encodes the title and available article or source-post text [2]. URLs and mentions are normalized, repeated whitespace is removed, and WordPiece tokenization is applied. Stop words are kept because function words and negation can carry contextual meaning.

For WordPiece tokens x1 to xn, the input sequence is:

$$\mathrm{T} = (\mathrm{[CLS]}, \mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_n, \mathrm{[SEP]}) \tag{1}$$

Here, [CLS] begins the sequence and [SEP] ends or separates segments. BERT maps the sequence to contextual states; the final [CLS] state is the document vector:

$$\mathrm{E}_t = \mathrm{BERT}(\mathrm{T})_{\mathrm{[CLS]}} \in \mathbb{R}^{768} \tag{2}$$

The 768-dimensional vector captures document-level semantics for fine-tuning, including contextual inconsistency, emotional framing, and unsupported certainty.

DOI: 10.54254/2755-2721/2026.BA36324

3.2. Social propagation graph modeling

For each news item, propagation is represented as a directed graph:

$$G = (V, E) \tag{3}$$

In Equation (3), nodes are source posts, replies, retweets, or users, and directed edges preserve the observed diffusion path. Node features may include timing, interaction type, account metadata, and compact text representations.

A GAT assigns a learned importance to each neighbor [3]:

$$\alpha_{ij} = \mathrm{softmax}_j(e_{ij}) \tag{4}$$

The softmax coefficient normalizes compatibility between a node and each neighbor using learned parameters. The next representation is the attention-weighted update:

$$h_i^{(l+1)} = \sigma(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}) \tag{5}$$

Multi-head attention stabilizes learning, and graph-level pooling yields a structural embedding. The branch captures concentrated amplification, short repost intervals, and influential bridge accounts; these patterns are evidence, not proof of malicious behavior.

3.3. Multimodal fusion and training objective

Text and graph embeddings are projected into the same d-dimensional space:

$$\widetilde{E}_t = W_t E_t + b_t, \qquad \widetilde{E}_s = W_s E_s + b_s \tag{6}$$

Learned projection matrices and biases align the channels. An attention gate then assigns each modality an item-specific contribution:

$$g = \mathrm{softmax}(W_g[\widetilde{E}_t \| \widetilde{E}_s] + b_g), \qquad F = g_t \widetilde{E}_t + g_s \widetilde{E}_s \tag{7}$$

The fused representation F can favor text for a sparse graph or propagation for short, ambiguous text, which is more flexible than fixed concatenation.

The channels are aligned by a consistency term:

$$\mathcal{L}_c = \| \widetilde{E}_t - \widetilde{E}_s \|_2^2 \tag{8}$$

Training combines classification and consistency losses:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_c \tag{9}$$

The coefficient lambda controls regularization. The objective discourages extreme disagreement without forcing identical channels. Dataset-specific output layers support binary FakeNewsNet labels and four-class Twitter15/16 labels. For LIAR, which has no native propagation graph, the graph branch is masked to evaluate missing-modality robustness.

DOI: 10.54254/2755-2721/2026.BA36324

4. Experimental setup

4.1. Datasets

The datasets differ in content, labels, and social context; Table 1 states which evaluations each can support.

Table 1. Dataset characteristics

Dataset	Main content	Social or structural information	Typical labels	Role in this study
FakeNewsNet	News articles, titles, and associated media	User engagements, temporal information, and social context for PolitiFact and GossipCop	Fake/real	Main full-fusion benchmark
LIAR	Short political statements with speaker and contextual metadata	No native retweet or reply propagation graph	Six truthfulness levels	Text-only and missing-modality test
Twitter15/16	Source tweets and conversation content	Propagation trees formed by replies and reposts	True, false, unverified, and non-rumor	Full-fusion and cross-domain evaluation

FakeNewsNet provides news plus social and temporal context and is the primary full-fusion benchmark [9]. LIAR contains about 12.8K labeled political statements but no native propagation graph [10]. Twitter15/16 provide source tweets and interaction trees for four-class rumor detection [4].

4.2. Preprocessing and graph construction

Preprocessing is fitted only on training data. Text is normalized and tokenized; long sequences retain the title, source claim, and opening content. Each propagation graph uses the source as root, preserves reply or repost direction, adds self-loops, removes duplicates, and records size, depth, and time span. Splits are made at event level to prevent related cascade branches from leaking across sets.

4.3. Implementation details

The PyTorch implementation fine-tunes BERT-base-uncased with AdamW, batch size 32, and 5-10 epochs. The GAT uses two heads and a 128-dimensional hidden state. Dropout, fusion dimension, and lambda are selected on validation Macro-F1, with early stopping.

4.4. Evaluation metrics

We report accuracy, Macro-F1, and binary-task AUC-ROC. Macro-F1 is emphasized for imbalanced and four-class data. Cross-domain robustness is measured by training on one Twitter dataset and testing on the other without target-domain fine-tuning.

DOI: 10.54254/2755-2721/2026.BA36324

4.5. Baselines

Baselines include TF-IDF/SVM, BERT-only, and FakeBERT for text [2, 7]; GAT-only and Bi-GCN for structure [3, 5]; and simple concatenation, GCAN, and FANG for fusion [6, 8]. Ablations remove consistency loss or replace adaptive attention with fixed concatenation. Comparisons are restricted to models receiving the information available in each dataset.

5. Results and discussion

5.1. Quantitative performance

On FakeNewsNet, the model reports 93.4% accuracy, 92.8% Macro-F1, and 0.97 AUC. BERT-only reaches 88.1% accuracy and GAT-only 86.3%, so fusion improves accuracy by 5.3 and 7.1 percentage points, respectively. Macro-F1 and AUC indicate gains beyond a single class or threshold.

The channels correct different errors. Text detects contradictions and manipulative framing but can be fooled by polished prose; graphs reveal coordinated amplification but can over-flag legitimate viral news. Fusion lets each channel qualify the other.

Across Twitter15/16, cross-domain accuracy declines by less than 2.8%, versus about 7-11% for unimodal baselines. Agreement between semantics and diffusion appears more stable than event-specific vocabulary.

Because LIAR lacks propagation graphs, it evaluates only the text pathway and masked-graph operation; no multimodal gain is claimed for that dataset.

5.2. Ablation study

Removing consistency loss lowers accuracy by 2.1-3.4 percentage points on full-fusion datasets. Without alignment, encoders can drift toward incompatible spaces and a branch can dominate through topical words or cascade size. Regularization reduces this domain-specific dependence.

Replacing adaptive fusion with concatenation reduces performance by about 1.5 points. Fixed concatenation cannot discount an incomplete channel, whereas attention lowers graph weight for sparse early cascades and can raise it for short or ambiguous text.

5.3. Interpretability

Interpretability operates at two levels. Token attribution identifies influential phrases, while GAT coefficients identify influential nodes and edges. A review interface can show the label, confidence, highest-weighted words, top nodes, and a compact cascade. These scores explain model behavior but do not prove that content is false or a user is malicious.

High attention to synchronized early reposts and phrases such as unsupported urgent claims can prioritize a case for review. Disagreement is also informative: suspicious diffusion around neutral language should prompt examination of the source, timing, and account context rather than automatic removal.

5.4. Discussion

The framework integrates what a news item says with how it spreads. Shared projections and consistency make the channels comparable, while attention varies their contribution by item. The

DOI: 10.54254/2755-2721/2026.BA36324

prediction is more auditable because analysts can inspect both linguistic and structural evidence.

Limitations remain. Graph evidence is weakest at publication time and interaction access depends on APIs, deletion, and privacy constraints; FakeNewsNet may require engagement data to be recollected [9]. Graphs can encode popularity or community structure rather than truth, creating fairness risks. BERT and GAT add computational cost, and attention is not a causal explanation. The model also does not detect manipulated images when diffusion appears ordinary.

Future work should add temporal and visual encoders, uncertainty calibration for incomplete graphs, model compression, fairness and adversarial testing, and privacy-preserving graph construction.

6. Conclusion

This study combines BERT semantics with GAT propagation structure through shared projections, adaptive fusion, and consistency regularization. Results and ablations show improvements over the supplied unimodal baselines, while word and node attributions support review. The system is intended to assist trained analysts, not make unchallengeable moderation decisions; deployment therefore requires temporal robustness, fairness evaluation, and privacy safeguards.

References

[1] Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*, Minneapolis, MN, 4171-4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[3] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada. https://openreview.net/forum?id=rJXMpikCZ

[4] Ma, J., Gao, W., & Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 708-717. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1066

[5] Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(1), 549-556. https://doi.org/10.1609/aaai.v34i01.5393

[6] Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 505-514. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.48

[7] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765-11788. https://doi.org/10.1007/s11042-020-10183-2

[8] Nguyen, V.-H., Sugiyama, K., Nakov, P., & Kan, M.-Y. (2020). FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, Virtual Event, Ireland, 1165-1174. Association for Computing Machinery. https://doi.org/10.1145/3340531.3412046

[9] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & 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. https://doi.org/10.1089/big.2020.0062

[10] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 422-426. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2067