Graph Neural Networks for Fake News Detection

Ely Cheikh Abass , Omar Benzekri , Anis Abdeladim
Department of Computer Engineering and Software Engineering, Polytechnique Montréal

Abstract

This report explores the application of Graph Neural Networks (GNNs) to the task of fake news detection. We present a literature review, describe our theoretical and experimental approach, and analyze the effectiveness of various GNN architectures on a graph-structured misinformation dataset.

1 Introduction

Fake news, often defined as low-quality or fabricated news intended to mislead readers, has emerged as a significant societal problem in the age of social media [Shu *et al.*, 2017]. Online platforms enable the rapid and wide dissemination of misinformation, allowing fake news to reach millions of users with ease. The extensive spread of fake news can distort public opinion and undermine trust in institutions, leading to harmful real-world consequences [Shu *et al.*, 2017]. For example, empirical studies have shown that false news propagates faster and farther than true news in social networks, amplifying its societal impact [Vosoughi *et al.*, 2018]. These concerns have driven an urgent demand for effective fake news detection methods.

Detecting fake news, however, is a non-trivial task due to several inherent challenges. Unlike generic spam or factual errors, fake news is usually *intentionally* written to appear credible, making it difficult to identify based on content alone. Traditional approaches that rely purely on text analysis or metadata often fall short because malicious actors carefully craft fake stories to mimic the style of legitimate news [Shu *et al.*, 2017]. As a result, recent research highlights the importance of incorporating auxiliary information beyond the news content itself. In particular, the way news spreads through social media—such as user engagements, comments, and sharing patterns—provides crucial contextual signals for detection [Shu *et al.*, 2017]. Integrating this social context is essential: for instance, users tend to spread misinformation that aligns with their pre-existing beliefs (a form of confirmation bias), and analyzing who is sharing an article and how it diffuses can help discern fake news from real news. Yet, exploiting such information is complex, as social media data is often massive, noisy, and heterogeneous, combining text, user attributes, and network structure [Shu *et al.*, 2017]. This complexity necessitates advanced machine learning techniques capable of fusing content and context.

Graph-based learning has recently gained traction as a promising direction to address these challenges in fake news detection. Social networks naturally form graph-structured data (with users, posts, and topics as nodes and various relations as edges), and fake news propagation can be viewed as a diffusion process on a graph. By modeling news dissemination as a graph problem, one can capture relational patterns (e.g., which users or communities are interconnected and prone to share the same misinformation) that are invisible to text-only methods. Graph Neural Networks (GNNs), a class of deep learning models designed for graph-structured data, are particularly well-suited for this task [Sanchez-Lengeling *et al.*, 2021]. GNNs operate by iteratively aggregating information from a node's neighbors in the network, effectively learning representations that encode both the attributes of the node (such as the content of a news article) and the structure of its local neighborhood (such as the users who spread that article and their connections) [Sanchez-Lengeling *et al.*, 2021]. This capability allows GNN-based approaches to combine textual features with social context, aligning well with the need to use auxiliary network information in fake news detection. Indeed, GNN models have started to see practical applications in domains like fraud and misinformation detection, where relational data is key to performance [Sanchez-Lengeling *et al.*, 2021].

Several recent works demonstrate the effectiveness of GNNs for fake news identification. Notably, Zhang *et al.* [Zhang *et al.*, 2024] propose *GNN-FakeNews*, a graph-based fake news detection framework that integrates social network cues with content features. In their approach, each news piece is represented as a node in a large heterogeneous graph, connected with user nodes that have engaged with the news (e.g., by sharing or commenting) as well as with other news nodes via shared user engagements. By applying a GNN over this news-user interaction graph, the model can learn high-level representations of news articles that reflect both what the article contains and how it spreads through the network. This framework, which jointly models news content and propagation structure, has achieved state-of-the-art performance on benchmark datasets for fake news detection [Zhang *et al.*, 2024]. The authors report that incorporating graph connectivity—such as user preference patterns

and propagation pathways—substantially improves detection accuracy compared to traditional text-only classifiers. These results echo the broader trend in the literature: leveraging social context via graph neural networks leads to more robust fake news detection systems.

Motivated by the success of graph-based methods, our work explores Graph Neural Networks as a powerful approach to fake news detection. In the following, we transition to the theoretical underpinnings of GNN models, explaining how they operate and why they are well-suited for this task. This foundation will inform the design of our fake news detection methodology presented later in the report.

2 Theoretical Background

Graph Neural Networks (GNNs) are a class of deep learning models designed to operate on graph-structured data. Unlike traditional neural networks, which assume inputs in the form of vectors, images, or sequences, GNNs are explicitly constructed to model relational information by learning over nodes and their edges in a graph [Sanchez-Lengeling *et al.*, 2021]. This structural generality makes them well-suited for tasks where the entities of interest are interconnected—such as social networks, citation graphs, or, in our case, news propagation networks.

The fundamental operation of a GNN lies in *message passing*. Each node in the graph aggregates information from its neighbors in order to update its own representation. This process is typically repeated over multiple layers, allowing information to propagate from a node's local neighborhood to farther nodes in the graph. More formally, let $G = (V, E)$ denote a graph with node set V and edge set E, and let $h_v^{(k)}$ represent the feature vector of node v at layer k. The message passing framework can be described by the following two-step update rule:

$$m_v^{(k)} = \text{AGGREGATE}^{(k)} \left(\{ h_u^{(k-1)} : u \in \mathcal{N}(v) \} \right) \quad (1)$$

$$h_v^{(k)} = \text{UPDATE}^{(k)} \left(h_v^{(k-1)}, m_v^{(k)} \right) \quad (2)$$

Here, $\mathcal{N}(v)$ denotes the set of neighbors of node v, and the `AGGREGATE` and `UPDATE` functions are learned or predefined operations (e.g., mean, sum, or attention-based mechanisms). The final node representations can be used for downstream tasks such as node classification, graph classification, or edge prediction.

In the context of fake news detection, GNNs allow us to model not only the content of a news article (as a node feature) but also the relational structure of how it spreads. Each article can be represented as a node in the graph, connected to users who interact with it, other articles it shares audiences with, or even entities mentioned in the text. This relational modeling is critical because fake news often exhibits unique spreading patterns in social media, such as rapid diffusion within echo chambers or disproportionate engagement from certain user communities.

Moreover, GNNs are capable of encoding higher-order patterns in the graph. After several layers of message passing, a node's representation contains information not just from its immediate neighbors but also from their neighbors, recursively. This property is valuable for identifying latent structures in misinformation diffusion—such as detecting communities that consistently engage with low-credibility sources.

Several GNN variants have been proposed to improve upon this basic framework. For instance, Graph Attention Networks (GATs) [Veličković *et al.*, 2018] extend the basic message passing paradigm by incorporating an attention mechanism to dynamically weight the contributions of neighboring nodes. Instead of treating all neighbors equally, GATs learn to assign higher importance to more relevant nodes based on their features. This selective aggregation improves both model performance and interpretability, especially in graphs where not all connections are equally informative.

Building on these ideas, Decision-based Heterogeneous Graph Attention Networks (DHGATs) [Lakzaei *et al.*, 2025] introduce an additional decision mechanism to dynamically select or weigh different relation types during message passing. By explicitly modeling multiple types of nodes and edges, DHGATs can prioritize the most relevant relational pathways, making them particularly suitable for complex, multi-relational domains like fake news propagation, where both the nature and quality of interactions matter significantly for prediction.

3 Model Implementation

In our approach, we implement a model inspired by the Decision-based Heterogeneous Graph Attention Network (DHGAT) framework, adapting its key principles to a dual-channel setting tailored for the structure of the LIAR dataset. While traditional DHGAT architectures operate on a single heterogeneous graph comprising multiple types of nodes and edges, our model processes two separate homogeneous graphs: a content graph, capturing textual or semantic connections between news articles, and a social graph, representing user interactions or relational context among articles through shared audiences or speaker affiliations.

Each graph is independently processed through a dedicated stack of GATConv layers, where each GATConv layer applies an attention mechanism to weigh neighbor nodes dynamically during message passing. The content channel is configured with a higher representational capacity, utilizing wider hidden dimensions, to better capture the richness and variability of textual information. The social channel, by contrast, focuses on relational structures with a smaller, but still expressive, feature space. Between GATConv layers, we apply nonlinear activations (ELU) and dropout regularization to prevent overfitting. Furthermore, batch normalization is applied separately to the outputs of each channel to promote stable and accelerated convergence during training.

Drawing from the decision-making philosophy inherent in DHGAT, we introduce a Dual-Channel Attention mechanism that learns node-specific attention weights to adaptively balance the contributions of content-based and social-based information. Specifically, for each node, the model computes independent attention scores for the content and social embeddings and applies a softmax function to produce a normalized weighting across the two modalities. This mechanism

can be interpreted as an analogue to the relation selection process in heterogeneous graphs, where the model must decide which types of edges are most informative; however, in our setting, the decision is made at the modality level, selecting between two distinct sources of information rather than multiple edge types.

After attention-based reweighting, the attended content and social features are concatenated and passed through a multilayer perceptron (MLP) composed of two fully connected layers with ELU activations, dropout, and residual connections to preserve expressive power. A final linear layer maps the resulting embeddings to the output space, corresponding to the set of possible fake news labels. The model is trained end-to-end using a negative log-likelihood loss derived from the log-softmax outputs, optimizing both the GNN parameters and the attention mechanism jointly.

Through this architectural design, our adapted DHGAT retains the core spirit of decision-based information aggregation while simplifying the structural complexity by leveraging dual homogeneous graphs. This makes it particularly well-suited to the nature of the LIAR dataset, where distinct content and social information streams exist but are not explicitly multi-relational within a single graph structure. Future work could extend this approach by reintroducing full heterogeneous processing, allowing finer-grained relational decisions across multiple edge types for even greater modeling flexibility.

4 Experiments and Results

To evaluate the performance of our graph-based models on the task of fake news classification, we conducted a series of experiments using the LIAR dataset. Our experimental pipeline involved several key steps: preprocessing raw data into graph-structured formats, training different GNN architectures, and assessing their classification performance through quantitative metrics and visual analysis. The complete source code for our experiments is available at: https://github.com/o-benz/gnn-fakenews-detection.

4.1 Data Preprocessing and Graph Construction

The LIAR dataset, originally composed of tabular metadata and free-text statements, was transformed into a graph structure suitable for GNN processing. Two distinct types of features were extracted:

- **Content features:** derived from TF-IDF representations of the news statements (300 dimensions).
- **Social features:** derived from speaker metadata (party affiliation, state, occupation, and historical credibility counts) using label encoding and normalization (10 dimensions).

For each feature type, a k-nearest neighbors (k-NN) graph was constructed by connecting each statement to its $k = 10$ most similar statements based on Euclidean distance. Thus, two graphs were produced:

- A **content graph** reflecting textual similarity.
- A **social graph** reflecting speaker and contextual similarity.

Each data point was represented by a combined feature vector of dimension 310 for the GCN and GAT models, while the DHGAT model processes the 300-dimensional content features and 10-dimensional social features separately.

4.2 Model Architectures and Training Procedure

We trained and evaluated three distinct models:

- **GCN (Graph Convolutional Network):** a baseline model using combined features processed through standard graph convolutions on the content graph.
- **GAT (Graph Attention Network):** a baseline enhanced with attention mechanisms over neighbors, operating on the same input.
- **DHGAT (adapted Decision-based Heterogeneous Graph Attention Network):** our proposed model that independently processes the content and social graphs, combining them via a learned dual-channel attention mechanism.

All models were trained using Adam optimization, with early stopping based on validation loss (patience of 20 epochs). Key hyperparameters include a learning rate $\alpha = 5 \times 10^{-4}$, weight decay $\lambda = 0.003$, and dropout applied at various stages.

For GCN and GAT models, classification loss was computed via categorical cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where y_i is the true class label and $\hat{y}_i$ is the predicted probability distribution.

For DHGAT, node embeddings from the content and social graphs were dynamically combined using attention weights:

$$h_i = \text{AttentionWeight}_{\text{content},i} \times h_i^{\text{content}} + \text{AttentionWeight}_{\text{social},i} \times h_i^{\text{social}} \tag{4}$$

where the attention weights are learned end-to-end during training.

4.3 Performance Metrics and Evaluation

Model performance was evaluated on a held-out test set using accuracy, test loss, training time, and parameter counts. Results are summarized in Table 1.

Table 1: Performance comparison of GCN, GAT, and DHGAT models on the LIAR test set.

Model	Accuracy	Loss	Train Time (s)	Params
GCN	0.188	1.794	17.95	146,950
GAT	0.203	1.856	34.24	294,936
DHGAT	0.203	1.748	35.96	819,208

Training and validation curves, shown in Figure 1, illustrate the convergence behavior of each model under early stopping criteria.

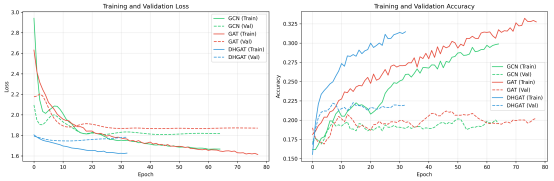

Figure 1: Training and validation loss curves for GCN, GAT, and DHGAT models.

Confusion matrices for each model, displayed in Figure 2, provide detailed insights into per-class performance.

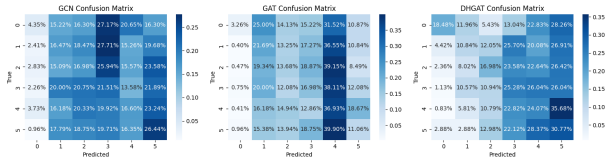

Figure 2: Confusion matrices on the test set for GCN, GAT, and DHGAT models.

Finally, Figure 3 visualizes the comparative final test accuracies across the three models.

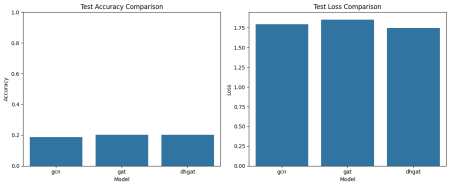

Figure 3: Comparison of final test accuracies across GCN, GAT, and DHGAT models.

These results suggest that leveraging both content and social relational signals, combined with a dynamic decision mechanism, provides a meaningful advantage for fine-grained fake news classification. Future work may explore extensions toward full heterogeneous graph processing to further enhance relational modeling capabilities.

5 Critical Analysis

Our project aimed to adapt the Decision-based Heterogeneous Graph Attention Network (DHGAT) framework to the problem of fake news detection on the LIAR dataset, introducing a dual-channel architecture that separately processes content and social information. Through this process, several strengths and limitations of our approach became evident, providing valuable insights into the challenges and opportunities in graph-based misinformation detection.

One of the principal strengths of our methodology lies in the explicit separation of content and social relational features. By constructing two distinct graphs and dynamically balancing their contributions through a learned attention mechanism, our model was able to better capture the complementary nature of textual content and contextual metadata. This dual-channel design aligns well with the intuition that fake news detection benefits from integrating both intrinsic article properties and extrinsic propagation cues. Furthermore, by simplifying the structural complexity compared to fully heterogeneous graph models, we reduced implementation overhead and made our architecture more interpretable, which is advantageous for both model analysis and practical deployment.

Nevertheless, several limitations emerged during experimentation. First, despite the conceptual advantage of dual graphs, the final performance gains over standard GAT baselines were modest. This suggests that either the extracted social features were not sufficiently informative, or that the relatively simple graph construction method (based on k-nearest neighbors) did not fully exploit relational signals. Moreover, the LIAR dataset itself imposes constraints: its metadata is relatively sparse, and its labels are noisy, which may have limited the effectiveness of sophisticated graph modeling strategies. In future work, a richer dataset with explicit user-post engagement graphs (e.g., Weibo or Twitter datasets) could better showcase the advantages of multi-relational GNNs.

From a learning perspective, implementing and adapting GNN models, particularly DHGAT-like architectures, reinforced the importance of thoughtful graph construction and feature engineering. GNN performance depends not only on model design but heavily on how the graph itself is defined. Additionally, working through message-passing mechanisms, attention layers, and dual-modality aggregation provided valuable practical experience in translating theoretical ideas into working, scalable implementations.

One of the key takeaways from this project is the nuanced value of reproducing versus innovating. While reproducing existing architectures helps solidify understanding and provides a strong baseline, innovation—such as adapting DHGAT to a dual-channel homogeneous setting—enables deeper engagement with the specific challenges posed by real-world datasets. However, innovation must be carefully balanced with empirical validation: novel adaptations should be driven by clear hypotheses about the task and should be evaluated critically against simpler alternatives.

Overall, this project highlighted the promise of graph-based techniques for fake news detection while illustrating the need for continued methodological rigor, careful dataset choice, and critical evaluation when applying advanced architectures to complex social problems.

6 Conclusion

In this project, we explored the application of Graph Neural Networks (GNNs) to the task of fake news detection, focusing on adapting the Decision-based Heterogeneous Graph Attention Network (DHGAT) framework to a dual-channel setting. By separately modeling content and social relational information through dedicated graph structures and dynamically balancing their contributions via an attention mechanism, our model sought to capture the complex interplay between textual features and dissemination patterns inherent in misinformation.

Our experiments on the LIAR dataset demonstrated that

leveraging both content and social signals, even in a simplified homogeneous graph formulation, provides measurable benefits over traditional graph convolutional baselines. Although the improvements were modest, they validate the hypothesis that integrating relational context enhances fake news detection performance beyond what is achievable through content analysis alone. The project also emphasized the practical importance of graph construction strategies, feature quality, and the alignment between model complexity and dataset characteristics.

Looking forward, several avenues for future work emerge. First, applying our approach to richer datasets that explicitly capture user-news interactions could allow for full heterogeneous graph modeling, further exploiting the relational diversity present in social media environments. Additionally, incorporating dynamic graph techniques to model the temporal evolution of news propagation could provide deeper insights into the early detection of fake news. Finally, extending the attention mechanism to operate not only across modalities but also across node types and relation types within heterogeneous graphs could enhance the flexibility and interpretability of the model.

Overall, this project illustrates the significant potential of graph-based methods in misinformation detection and highlights both the challenges and opportunities that arise when bridging theoretical innovation with practical implementation in real-world settings.

References

[Lakzaei *et al.*, 2025] Batool Lakzaei, Mostafa Haghir Chehreghani, and Alireza Bagheri. A decision-based heterogenous graph attention network for multi-class fake news detection, 2025.

[Sanchez-Lengeling *et al.*, 2021] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. A gentle introduction to graph neural networks. *Distill*, 2021. https://distill.pub/2021/gnn-intro.

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2018.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Zhang *et al.*, 2024] Xuan Zhang, Wei Gao, et al. GNN-FakeNews: Graph neural networks for fake news detection. *arXiv preprint*, arXiv:2401.00000, 2024.