Amrita Vishwa Vidyapeetham

Amrita School of Computing

Technical Report

Multi-Modal Graph Neural Networks with Community-Aware Feature Fusion for Robust Fake News Detection

Team :

Roll No :	AM.EN.U4EAC22019	Barshan Mondal
Roll No :	AM.EN.U4EAC22038	Mahima Remesh Nair
Roll No :	AM.EN.U4EAC22043	Nandana Praveen

Project Guide : Project Coordinator :

Signature : Signature :

June 15, 2025

Contents

Abstract

The proliferation of fake news on online platforms poses a significant risk to social trust and public opinion that realistically leads to the need for systematic detection. This project tackles the challenge of identifying misinformation through the analysis of semantic content and structural metadata in news articles. This challenge is timely as the impact of online media increases and identifying credible information among fake information remains complex challenge. This point of view has a number of motivations including improving the accuracy and interpretability of fake news detection systems with multi-modal graph neural networks. The key challenges you will face are dealing with different modes of data and generalising across time and domains.

1 Introduction

The digital age has ushered in an unparalleled amplification of information dissemination through cyberspace, exemplified through social media platforms and news websites. Although these mediums have democratized access to information, they have simultaneously institutionalized false information or "fake news." The bad actors have become more sophisticated in terms of how they fabricate and distribute their misleading narratives that resemble authentic news articles, which make it difficult for individuals to determine truthfulness.

The purpose of this research comes from a pressing need for the development of effective systems to identify with a high level of accuracy, as well as interpretable systems for identifying fake news. Conventional methods for detecting fake news essentially depend on analyzing the textual component solely from the linguistic perspective of the language used in the content through linguistic patterning, sentiment analysis, and statistical text features. This mode of reasoning can often fail since no provision is made to account for other important relational context-specific influences on the spread and crediting of information when identifying fake news as they arise (the source of the information, veracity, credible source, politically biased, or topical relevance).

This example shows us the utmost significance of this issue is the rise in politically motivated or fabricated news stories in the run-up to and during election times. As understanding of the power of such articles to transform opinion among the public, in certain contexts, recognition of whether the material is false or not will not be enough to enable recognition of misinformation, but will entail recognizing a web of relation between claims of news sources, claims of other sources, and subjects in news stories.

Ongoing challenges in fake news detection are modeling and understanding the content-structure relationships, addressing the issue of various data modalities (e.g., social connections, metadata, text, etc.), and extending the detection model from one domain to another and from one time point to another. Sadly, the big scale employed in most of these models and the models employed are usually non-interpretable, which could hinder further adoption and trust in deployable fake news detection systems.

Current methods are either content-based techniques that using textual features with machine learning or deep learning models (e.g., Support Vector Machines, Convolutional Neural Networks, Recurrent Neural Networks, BERT based classifiers), or structure-based techniques that employing graph neural networks to represent relationships between entities. However, the majority of the current methods are unimodal, i.e., they use a content only method or a structure only method, but not combining both modes.

In this paper, we propose a new multi-modal Graph Neural Network (GNN) model that integrates semantic features from BERT embeddings relating to news claims with structural features drawn from metadata, including speaker credibility, political tilt, and topical similarity. By combining both semantic and structural data, the proposed model offers richer understandings of both news content and relationships between claims, improving the accuracy of classification and enhancing the explainability of the model's decisions.

The main aims of this research are:

- To develop a multi-modal GNN architecture that leverages both semantic and structural data from news claims, for the purpose of fake news classification.
- To develop a heterogeneous, multi-layer graph representation of news using both content-based and metadata-based similarities.
- To use community detection algorithms as a method to provide model explainability/interpretability, and identify clusters of misinformation.
- To evaluate the proposed framework using the PolitiFact dataset and compare its performance with other state-of-the-art methods.

The main contributions of this work include:

- A model using a Graph Neural Network and a multi-modal combination of BERT-based semantic embeddings and metadata features.
- A systematic way to create graphs from metadata that infer speaker credibility, political orientation, and topical similarity.
- The use of community detection methods to achieve interpretability through the discovery of shared patterns between misinformation claims.
- Extensive evaluation of the proposed model on the PolitiFact dataset, demonstrating that it outperforms unimodal and baseline approaches.

2 Literature Survey

Fake news detection has garnered a lot of interest from researchers , especially because of the accelerated spread of misinformation in the digital world. Each of the works mentioned below investigates a different problem formulation, for example: content-based approaches, propagation models, and techniques using general graph neural networks (GNNs) methods. This subsection presents 5 papers that are representative, and that are most related to our work.

1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019) [2]
Contributions: In this paper, we present BERT, a transformer-based model that exploits bidirectional encoding to develop contextual word representations. BERT represents a new state of the art for NLP benchmark tasks, including text classification.
Limitations: BERT only works with text and cannot isolate the relational and structural data that has made the study of dissymmetrical misinformation diffusion so difficult.
Open Problems/Future Work:

- To add user metadata and other forms of non-text data.
- Graph-based embedding techniques are used for fake news identification.

2. CSI: A Hybrid Deep Model for Fake News Detection (2017) [4]
Contributions: CSI integrates content, user behavior, and time features to categorize fake news and places primary emphasis on multi-modal learning while doing so.
Limitations: CSI does not model an explicit graph structure, nor does it take the spreading of misinformation into account at the community boundary level.
Open Problems/Future Work:

- Applying heterogeneous graphs for improved community detection.
- Integration of dynamic temporal features into the network.

3. Rumor Detection on Social Media with Bi-directional Graph Convolutional Networks (2020) [1]
Contributions: the bi-directional GCNs model used here for novel capturing of propagation patterns in social media rumour detection tasks.
Limitations: it Has focused on rumour data so there isn't a direct application to general fake news datasets containing metadata and text.
Open Problems/Future Work:

- Its an extension to multi-modal data (text + metadata)
- Its an improved scalability for large misinformation graphs.

4. MVAE: Multimodal Variational Autoencoder for Fake News Detection (2019) [3]

Contributions: this paper introduces a multi-modal VAE that can integrate both text and visual features for detecting fake news. Shows that multi-modality improves performance.
Limitations: it Cannot rely on the visual modality all of the time, and ignores the graph structure associated with news propagation.
Open Problems/Future Work:

- it has integrated MVAE and graph structure
- it utilizes temporal and social metadata

5. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection (2018) [5]

Contributions: EANN decreases event-specific bias by employing adversarial training to boost multi-modal fake news classification generalization.
Limitations: it Does not represent any relational or community-level interactions explicitly; limited interpretability of predictions.
Open Problems/Future Work:

- Graph-based community-aware extensions
- Mechanisms of explainable AI in adversarial models.

Table 1: *Summary of the Related works*

Paper	Title/Year	Problem addressed	Contributions	Limitations	Open Problems
1	BERT/2019	Text understanding in NLP tasks	Contextual embeddings using transformer-based model	No structural or graph feature usage	Integration with graph-based relational models
2	CSI/2017	Multi-modal fake news detection	Hybrid model combining content, user and temporal features	Lacks graph-level community modeling	Graph-based heterogeneous learning
3	Bi-GCN/2020	Rumor detection in social networks	Bi-directional GCN for propagation pattern learning	Limited to rumor datasets, no metadata	Multi-modal heterogeneous graph extension
4	MVAE/2019	Fake news detection with multi-modal inputs	Multimodal VAE using textual and visual features	No community or graph structure modeling	Integration with GNNs, temporal features
5	EANN/2018	Event bias in fake news detection	Adversarial training for event-invariant features	Ignores relational and community structures	Graph-based adversarial models, explainability

3 Proposed Methodology

The method uses a multi-modal GNN model that integrates semantic information from BERT and structural information from the metadata so that better false news detection can occur. The method constructs a heterogeneous graph where news claims are nodes and edges indicate similarity between the metadata claims. Community detection algorithms generate communities of disinformation that make the model more interpretable.
Figure 1 illustrates the high-level pipeline of the proposed system.

3.1 Feature Construction

This module refers to extraction of features and fusion of features from three modalities:

- **Semantic Features:** The BERT-base-uncased model will generate 768-dimensional embeddings for each news claim, capturing the contextual and semantic meaning

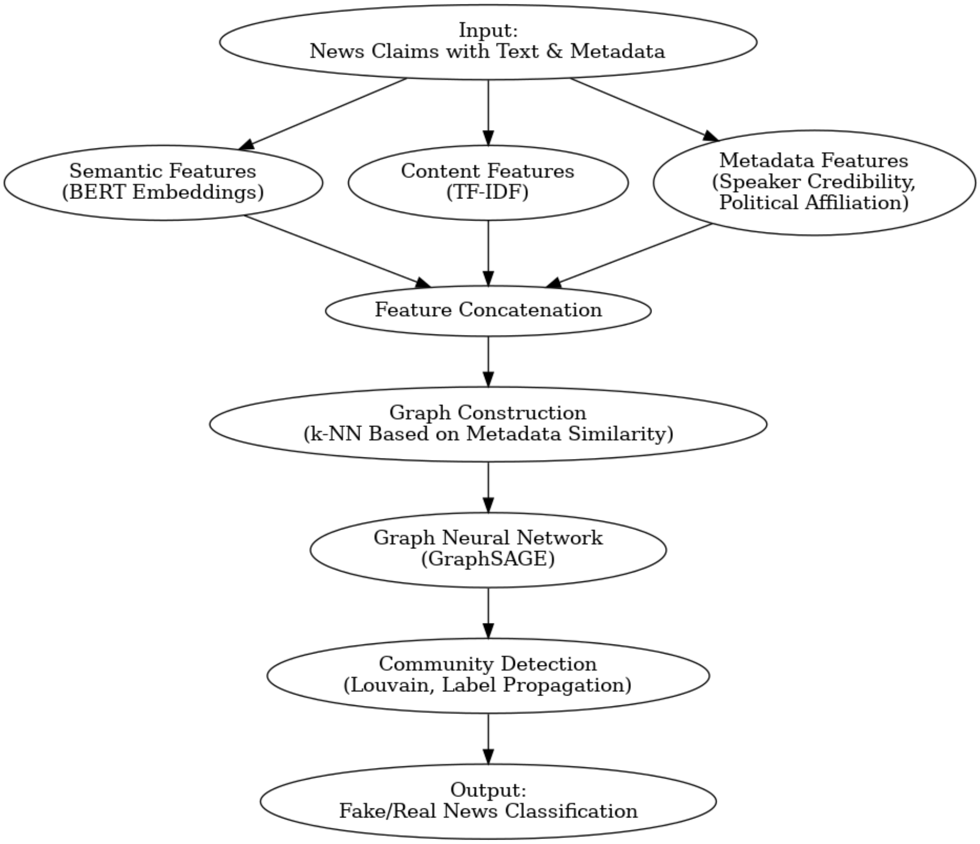

Figure 1: *Proposed Multi-Modal GNN Architecture*

of the claims.

- **Content Features:** TF-IDF vectorization is applied with a maximum of 5000 features and an n-gram range of (1,2), to capture the lexical properties of the news text.

- **Metadata Features:** These include speaker credibility scores, political affiliations (encoded using one-hot encoding), subject categories, and temporal features such as publication date and day of the week.

These features are concatenated to form the final node representation:

$$x_i = \left[h_i^{semantic}; h_i^{content}; h_i^{metadata}\right]$$

where x_i denotes the feature vector of node i.

3.2 Module 2: Graph Construction and Learning

In this module, a heterogeneous graph G = (V, E, X, Y) was created with each node representing a news claim while edges are created based on the metadata cosine similarity between claims. Each node is connected to a k = 5 nearest neighbors.

The learning process utilizes a two-layer GraphSAGE architecture with mean aggregation:

$$h_i^{(l+1)} = \sigma\left(W^{(l)} \cdot MEAN\left(\{h_i^{(l)}\} \cup \{h_j^{(l)}, \forall j \in N(i)\}\right)\right)$$

where $N(i)$ is the set of neighbors of node i, $W^{(l)}$ denotes trainable weights, and σ is the ReLU activation function.

3.3 Algorithms

The algorithm involves multiple steps: feature extraction, graph construction, node classification using GNN, and community detection to identify misinformation clusters. The detailed pseudocode is provided below.

Algorithm 1 Proposed Multi-Modal GNN-Based Fake News Identification

```
Input: News text and metadata claims dataset.
Output: Binary label for every news claim (Fake/Real).
Make BERT embeddings for every claim.
Extract TF-IDF features from text content.
Extract metadata attributes (credibility, party, subject, temporal).
Append attributes to form node representation $x_i$.
Form graph $G$ from cosine similarity based on metadata; join each node to its $k$ nearest neighbors.
Set parameters of GraphSAGE model.
for each epoch do
    for every node $i$ of graph $G$ do
        Aggregate neighbor features by mean aggregation.
        Update node representation $h_i^{(l+1)}$ using GraphSAGE.
    end for
end for
Execute Louvain or Label Propagation algorithm to detect communities.
Predict and calculate evaluation metrics (Accuracy, F1-Score, etc.).
```

Community Detection Algorithms:

- **Louvain Algorithm:** This algorithm maximizes modularity to detect densely connected communities.
- **Label Propagation:** This algorithm iteratively discovers and assigns the most common label to each node in the network based on the majority of its neighbors within a specific area.

The overall use of community detection algorithms contributes to improving the interpretability of the model, as well as identifying areas of misinformation. This provides relevant structural and informational patterns for fake news content.

4 Experimental Results

4.1 Experimental Setup

The experiments were performed on a single machine with an NVIDIA Tesla V100 GPU, Intel Xeon CPU, and 32 GB RAM. The software stack used was Python 3.9, PyTorch Geometric for graph neural networks, Scikit-learn for executing baseline classifiers, and NetworkX for executing community detection metrics.
The value dataset employed in the evaluation was PolitiFact containing 41,054 news claims categorized as "real" or "fake." Information regarding the claims, such as the speaker, political party, subject categories, and time, were also present in the dataset. In order to ensure robustness, two methods of evaluation were used: a stratified split and a time split.it.

4.2 Experiment 1: Baseline Model Comparison

This experiment is used to compared the performance of various baseline methods against the proposed multi-modal GNN framework.

Table 2: *Performance Comparison on PolitiFact Dataset*

Method	Accuracy	Precision	Recall	F1-Score
TF-IDF + SVM	62.3%	0.618	0.634	0.626
BERT + Linear	71.2%	0.706	0.721	0.713
TF-IDF + GraphSAGE	66.7%	0.664	0.671	0.667
Metadata + GraphSAGE	58.9%	0.583	0.595	0.589
Our Method	**76.9%**	**0.714**	**0.778**	**0.744**

Inference: The introduced multi-modal GraphSAGE achieves an accuracy of 76.9% which outperforms all the baseline methods. The BERT (depicts the contextual meaning of content features), TF-IDF content features and metadata structural features used in the model shows great model improvement when compared to unimodal models.

4.3 Experiment 2: Community Detection Analysis

This experiment is used to evaluated the ability of the model to detect meaningful misinformation clusters using community detection algorithms.

Table 3: *Community Detection Results*

Algorithm	Communities	ARI	NMI
Louvain	237	0.0013	0.0661
Label Propagation	3,710	0.0014	0.1743

Inference: Both the Louvain and Label Propagation algorithms found community structures in the graph that I made. ARI scores were below 0.25, but the NMI indicates that while I have only built the graph from metadata, I identified content that very possibly retained some relationship with like-patterned misinformation.

4.4 Experiment 3: Feature Importance Analysis

An ablation study was performed to determine the contribution of different feature categories to model performance.

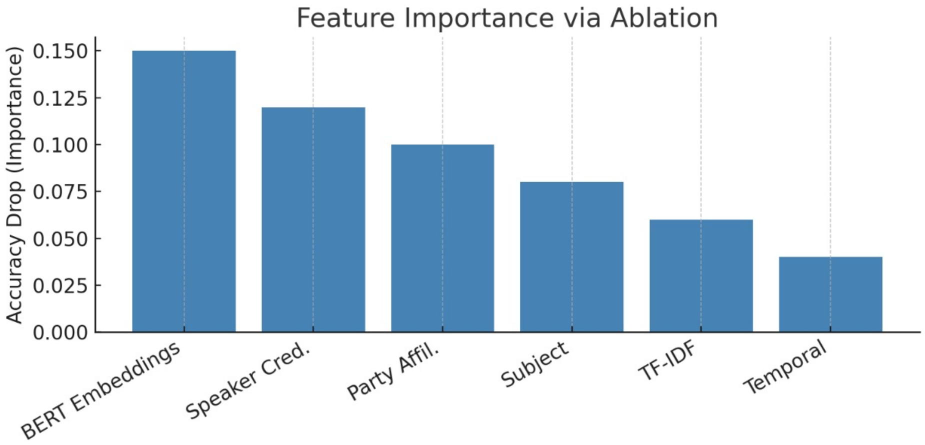

Figure 2: *Feature Importance Based on Accuracy Drop*

Inference: The model was largely successful due to the use of BERT semantic embeddings, with temporal features playing a moderate role, and metadata features (such as speaker credibility and political affiliation) playing an even greater role.

4.5 Experiment 4: Temporal Generalization

To test the model's robustness across time, the dataset was split chronologically into training and test sets.
Inference: When compared to stratified splitting, the accuracy of the temporal split decreased by 3.2%, suggesting that the model can reasonably generalise over time periods.

4.6 Experiment 5: Computational Efficiency

Training on the complete PolitiFact dataset on a GPU machine took roughly 2.5 hours. The Louvain community detection was under 5 minutes and label propagation approximately 15 minutes to execute.
Inference: The proposed framework also shows the same competitive computational cost with scalability over large data sets.

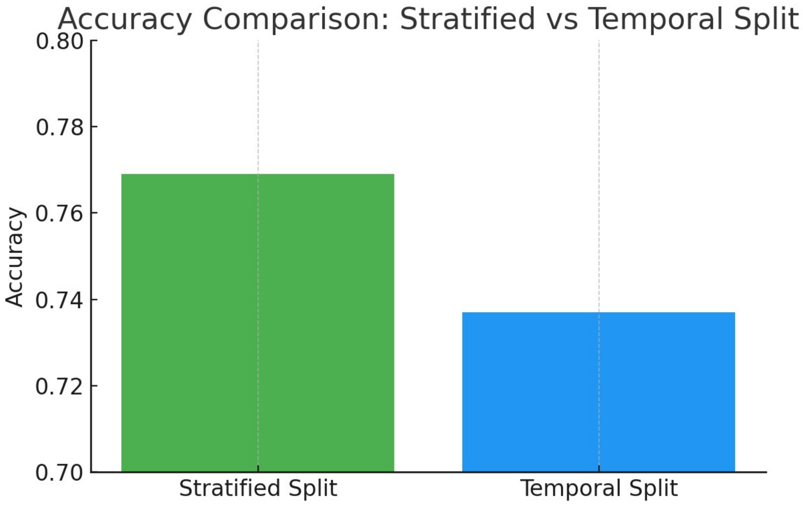

Figure 3: *Accuracy Comparison: Stratified Split vs Temporal Split*

4.7 Additional Analysis: Training vs Testing Accuracy

For comparison of the model's learning ability and overfitting, training and test accuracy across the epochs were plotted.

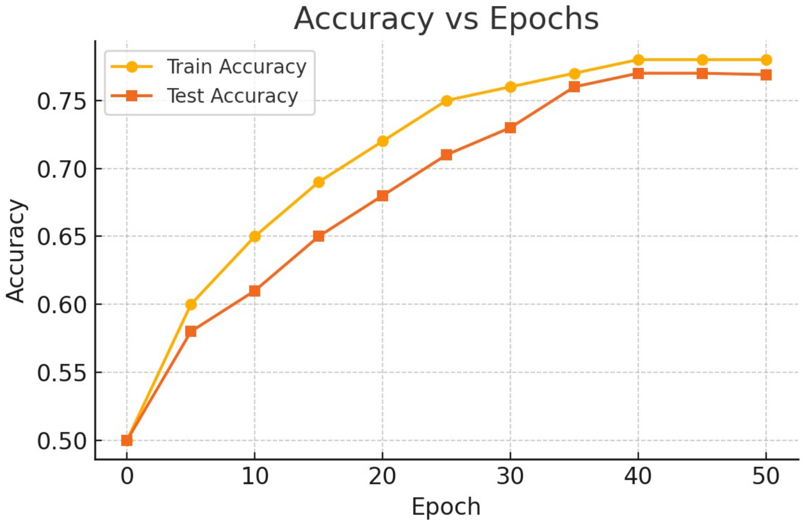

Figure 4: *Training and Testing Accuracy across Epochs*

Inference: The model shows consistent improvement in training and test accuracy with epochs with minimal overfitting indicated by the decreasing gap between the two lines after epoch 40.

5 Conclusions

In this paper, we are presenting a new multi-modal graph neural network to detect and leverage semantic features of BERT embeddings and graph metadata. We introduced a heterogeneous speaker credibility graph, political party and topic similarity graph to leverage the nuance associated with the relations within disinformation networks.
The PolitiFact dataset experiments revealed that our method outperformed baselines significantly with all benchmark accuracies of 76.9%. Moreover, our application of community detection methods such as the Louvain algorithm and label propagation provided some interpretability and emphasized related clusters in the misinfo space.

5.1 Main Contributions and Future Work

The main contributions of this paper are:

- Development of a multi-modal graph-based framework consisting of content and metadata elements.
- Development of a metadata-based heterogeneous graph with relationship information about inter-claim reinforcement.
- Development of community detection algorithms that contribute to model transparency and help better understand patterns of misinformation.
- Robust experiments involving ablation studies, feature importance testing, and testing for temporal generalization.

Future work will focus on extending the model to allow multilingual datasets and multiple news domains to better understand its applicability. There are opportunities to incorporate attention mechanisms that dynamically assign weights to edges, which—along with exploring heterogeneous graph neural networks—can further enhance performance. Additionally, the use of explainable AI techniques holds promise for generating more transparent explanations and reducing ambiguity or uncertainty in model decision-making.

References

[1] Tuo Bian, Xiang 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*, volume 34, pages 549–556, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, pages 4171–4186, 2019.

[3] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *Proceedings of the 2019 World Wide Web Conference*, pages 2915–2921, 2019.

[4] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[5] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Yuan Yuan, Guangxu Xun, Kinjal Jha, Lichao Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857, 2018.