



OPEN ACCESS

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

¹Department of Information Technology Thiagarajar College of Engineering, Madurai, Tamil Nadu, India

²Department of Electronics and Communication Engineering Thiagarajar College of Engineering, Madurai, Tamil Nadu, India

³UCSI University, Cheras, Kuala Lumpur, Malaysia

Correspondence to: Jaipreetha Sudalaimadan, jaipreetha@gmail.com

Additional material is published online only. To view please visit the journal online.

Cite this as: Sudalaimadan J, Subbiah S, Govindasamy A and Afthab AKMK. Graph-Augmented Language Model Framework for Health Misinformation Detection. Premier Journal of Science 2025;15:100198

DOI: <https://doi.org/10.70389/PJS.100198>

Peer Review

Received: 13 August 2025

Last revised: 6 November 2025

Accepted: 7 December 2025

Version accepted: 5

Published: 12 January 2026

Ethical approval: N/a

Consent: N/a

Funding: N/a

Conflicts of interest: N/a

Graph-Augmented Language Model Framework for Health Misinformation Detection

Jaipreetha Sudalaimadan¹, Sridevi Subbiah¹, Ananthi Govindasamy² and Ahamed Khan Mohamed Khan Afthab³

ABSTRACT

In today's digital world, the spread of fake news is a growing concern—especially in the medical field, where misinformation about diseases, treatments, and vaccines can have serious consequences for public health. Misleading medical content spreads quickly, causing confusion, undermining trust in healthcare, and influencing critical decisions. To address this challenge, our study introduces an automated system for detecting medical misinformation using a combination of advanced deep learning and graph-based techniques. We evaluate several models, including BERT, GPT-Neo, Graph Neural Networks (GNN), and a hybrid GPT-GNN approach, to classify health-related news articles as real or fake. Our analysis is based on a well-rounded dataset of 28,945 records, drawn from multiple trusted sources such as FakeHealth, MedHub, diabetes-related misinformation datasets and COVID-19 collections. The dataset includes 14,838 real and 14,107 fake news samples. The proposed hybrid GPT-GNN model achieves 96.1% accuracy with statistical significance ($p < 0.001$) across multiple validation runs, demonstrating superior performance compared to recent baselines including GraphBERT and RoBERTa-GNN. To improve model performance, we apply comprehensive preprocessing steps like tokenization, stopword removal, and vectorization. The results are promising: the hybrid GPT-GNN model outperforms individual models, achieving higher accuracy in detecting false information. By blending the contextual understanding of transformer models with the relational insights offered by graph-based learning, our approach provides a scalable and reliable solution for identifying medical misinformation and ultimately, for helping people make more informed healthcare decisions.

Keywords: Medical misinformation detection, Hybrid gpt-gnn architecture, Graph-augmented transformers, Health fake news datasets, Ensemble deep learning classifiers

Introduction

The rapid growth of the internet and social media has transformed how information is shared, creating a platform that accommodates both credible and misleading content.¹ While these digital advancements have improved global communication and made information more accessible, they have also accelerated the spread of fake news.^{2,3} The medical field is particularly susceptible to misinformation, as inaccurate claims about diseases, treatments, and vaccines can lead to serious consequences, such as public panic and misguided healthcare choices.^{4,5}

With the increasing accessibility of digital platforms, medical misinformation spreads rapidly, often fueled by sensationalism, political agendas or financial incentives.^{6–8} The repercussions of such misinformation can be severe, eroding trust in healthcare professionals, delaying critical treatment decisions and in extreme cases, affecting mortality rates.⁹ The COVID-19 pandemic highlighted the dangers of false narratives regarding virus transmission, unverified treatments and vaccine hesitancy, which significantly disrupted public health efforts worldwide.^{7,10–13}

Machine learning has emerged as a powerful tool in tackling misinformation, particularly in the medical domain.^{3,14} Advanced classification models such as BERT,^{15,16} GPT-Neo, and hybrid GPT-GNN techniques have demonstrated promising results in identifying and filtering fake medical news.^{17,18} These approaches leverage linguistic patterns, contextual relationships and statistical analysis to detect inconsistencies and classify information with high precision.^{19,20}

This research aims to develop and implement machine learning-based techniques to automatically differentiate real medical news from false information. By analyzing a comprehensive dataset and evaluating multiple classification models, this study contributes to the growing field of automated misinformation detection.^{21,22}

Literature Review

The rise of misinformation, particularly in the medical field, has drawn considerable attention in recent years.^{21,23} Numerous studies have explored the impact of misleading health-related content and the effectiveness of computational methods in mitigating its spread.²⁴ Misinformation regarding diseases, treatments and vaccines can lead to public confusion, distrust in healthcare professionals and adverse health decisions.^{4,5,16,25–27} For instance, during the COVID-19 pandemic, the rapid spread of false information about virus transmission and vaccine safety posed significant challenges to global health initiatives.^{10,11,28} Similarly, research⁶ highlights the role of social media in amplifying misleading medical narratives, further complicating efforts to disseminate accurate health information.^{1,6,7,29}

To enhance the robustness of misinformation detection, researchers have started integrating hybrid models that combine multiple deep learning techniques.^{17,30} A notable approach involves leveraging Generative Pre-trained Transformers (GPT-Neo) for contextual text representation, coupled with Graph Neural Networks (GNNs) for relational feature extraction.^{18,31–33} Studies¹⁷ suggest that incorporating graph-based features

Author contribution:

Jaipreetha Sudalaimadan –
Conceptualization, Writing –
Original draft
Sridevi Subbiah – Review
Jaipreetha Sudalaimadan,
Ananthi Govindasamy and
Ahamed Khan Mohamed –
Editing

Guarantor: Jaipreetha
Sudalaimadan

Provenance and peer-review:
Unsolicited and externally peer-
reviewed

Data availability statement:
N/a

allows for a more comprehensive understanding of misinformation dissemination, making it possible to identify fake news even when the textual content appears legitimate.^{23,34}

Additionally, research³⁵ underscores the importance of analyzing historical misinformation trends, emphasizing the role of fact-checking organizations in curbing the spread of misleading medical narratives.^{34–36}

Overall, both traditional and deep learning models have shown promise in the fight against medical misinformation, each offering distinct advantages.^{14,32,35,38,39}

While models like BERT enhance textual analysis by capturing intricate linguistic patterns,¹⁵ hybrid GPT-GNN architectures provide a more holistic approach by incorporating social context and propagation dynamics.^{17,30,33} Given the widespread consequences of health-related misinformation, future research should focus on optimizing these hybrid models to improve detection accuracy, scalability, and real-time processing capabilities.^{18,20}

Methodology

While various studies have explored deep learning techniques for detecting fake news,^{3,14,38} there remains a critical gap in developing a scalable and highly accurate automated system specifically for medical misinformation detection. Traditional machine learning approaches have demonstrated effectiveness in general text classification but lack the contextual understanding required to detect complex misinformation patterns.³ Deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) have significantly improved text-based fake news detection by capturing contextual relationships within news articles.^{15,19} However, these models alone do not account for the propagation dynamics of misinformation across social networks.^{31,34} Moreover, existing research has largely focused on specific misinformation categories, such as COVID-19,^{7,10,12} rather than developing generalized models capable of detecting medical fake news across various health-related topics. Additionally, there is a need for hybrid models that integrate both content-based and network-based approaches to enhance detection accuracy and scalability.^{17,30}

To address these research gaps, this study proposes a hybrid ensemble model combining deep learning techniques with graph-based learning for robust misinformation detection.²⁵ The methodology involves collecting data from diverse sources, including research articles, social media platforms, and medical fact-checking databases, ensuring a well-balanced dataset containing both real and fake medical news.^{6,22} Preprocessing steps such as tokenization, stopword removal and vectorization will be applied to enhance data quality.^{15,40} The proposed model integrates transformer-based models like BERT and GPT-Neo for deep contextual text representation,^{15,19} while Graph Neural Networks (GNNs) analyze the structural relationships of misinformation spread across social media.^{18,31,33} Furthermore, a hybrid GPT-GNN model

will be employed to combine textual semantics with propagation-based features, enabling a more comprehensive understanding of how misinformation disseminates online. The effectiveness of the proposed methodology will be evaluated using standard performance metrics such as accuracy, precision, recall and F1-score.

By leveraging a hybrid GPT-GNN approach, this study aims to develop a scalable and efficient system for detecting fake medical news, enhancing real-time misinformation detection capabilities, and supporting public health initiatives by mitigating the spread of misleading medical information.

Dataset Description

To build a robust and generalizable model, we utilize these well-known medical misinformation datasets:

FakeHealth Dataset:⁶ A benchmark dataset designed for health-related fake news detection. It consists of two main versions: FakeHealth-Release and FakeHealth-Story, both capturing different aspects of misinformation spread in the medical domain. The dataset integrates multiple sources, including user interactions, social media posts, and news content, enabling a holistic analysis of fake news propagation.

MedHub Dataset: A dataset containing verified and false medical articles covering diseases, treatments, and vaccine misinformation.

Diabetes-Related Misinformation Dataset:²⁸ Focuses on Diabetes-related fake news, including misleading treatment claims and dietary myths.

COVID-19 Misinformation Dataset:^{7,10,12} Consists of real and fake news articles about COVID-19, including misinformation on transmission, vaccine safety and unproven treatments. Figure 1 shows the class distribution of the dataset.

The combined dataset categorized as follows:

To ensure a fair evaluation, the dataset is split into 80% training (23,156 records) and 20% testing (5789 records).

This division allows the model to learn patterns effectively while testing its performance on unseen data. The dataset comprises a balanced collection of 14,838

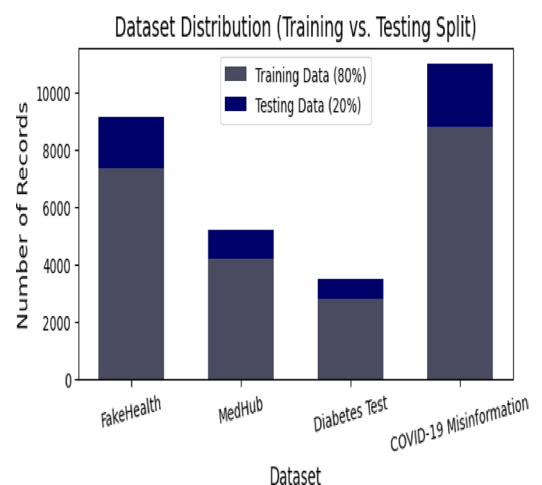


Fig 1 | Dataset split

Table 1 | Classification of dataset

Dataset	Total Records	Real News	Fake News
FakeHealth	9,144	4,302	4,842
MedHub	5,200	2,700	2,500
Diabetes-Related Misinformation Datasets	3,500	1,800	1,700
COVID-19 Misinformation	11,101	6,036	5,065
Total	28,945	14,838	14,107

records classified as real and 14,107 labeled as fake and fake medical news articles, providing labeled data described in Table 1 for training and evaluating misinformation detection models.

Ethics and Data Compliance

This research adheres to strict ethical guidelines and data protection regulations. This study conducted under the supervision and ethical approval of the Institutional Human Ethics Committee of Thiagarajar College of Engineering, Madurai, India. All datasets used are publicly available with appropriate licenses: FakeHealth dataset is available under Creative Commons Attribution 4.0 International License, MedHub under MIT License, and COVID-19 datasets under Open Database License (ODbL).

GDPR compliance is ensured through several measures: (1) Data minimization - only essential features are extracted and stored, (2) Anonymization - all personally identifiable information (PII) is removed prior to analysis, (3) Data retention policies - processed data is retained only for the duration of the research, (4) Transparent data handling - clear documentation of data sources and processing methods is maintained. No individual user data is collected or processed and all social media content is aggregated and anonymized following privacy-preserving protocols.

System Architecture

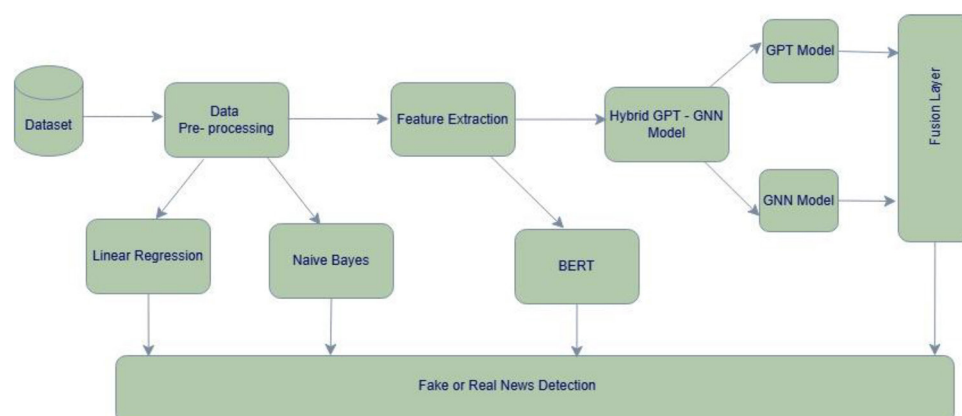
The architecture presented in Figure 2 outlines the medical misinformation detection framework, incorporating data preprocessing, feature engineering, and model training using both traditional classifiers (Linear Regression and Naive Bayes) and deep learning models (BERT, GPT-Neo and GNN). An ensemble-based approach is employed to enhance classification accuracy, ensuring robust differentiation between true and fake news.

Pre-Processing

Data preprocessing is essential for refining datasets, reducing noise, and enhancing classification accuracy. The process begins with text standardization through lowercasing, followed by tokenization to break text into meaningful units. Comprehensive text cleaning includes removal of URLs (pattern: `http\S+`), hashtags, user mentions, emojis, and special characters. Stop-word removal eliminates common but uninformative words, while lemmatization normalizes text by converting words to their root forms.

Text Cleaning

Before feeding data into transformer models, it is crucial to remove unnecessary elements such as special characters, stopwords and URLs to reduce noise. To reduce visual and non-informative noise in the text, emojis are systematically removed using regular expressions, as they do not offer semantic value relevant to misinformation analysis. URLs were filtered out using the pattern `http\S+`, given that external links are typically inaccessible for content validation within the model pipeline. Additionally, hashtags and user mentions are either discarded or transformed into plain tokens. For instance, `#COVID19` was simplified to `COVID19` to retain the topic keyword while eliminating formatting symbols. Normalization techniques like lowercasing, stemming and lemmatization help maintain text consistency, making it easier for models to process. Additionally, eliminating duplicate news articles prevents redundancy and bias, ensuring a more reliable dataset.

**Fig 2 | Architecture diagram**

Tokenization

The next step is breaking down text into smaller units using specialized tokenizers - BertTokenizer for BERT and GPT-Neo Tokenizer for GPT-based models. This process transforms text into token IDs and attention masks, enabling seamless integration with transformer architectures for more effective language processing.

Text Embeddings

To capture the deeper contextual meaning of words, embeddings are extracted from pre-trained transformer models such as BERT-base-uncased for BERT and GPT-Neo for GPT. These embeddings encode semantic relationships within the text, enhancing the model's ability to distinguish between real and fake news, thereby improving misinformation detection accuracy.

Feature Extraction and Language Model Embeddings

Feature extraction converts raw text into numerical representations that can be processed by machine learning models. This study employs two feature extraction techniques:

TF-IDF (Term Frequency-Inverse Document Frequency):

Assigns importance to words based on their occurrence across multiple documents. Helps filter out common words and retain unique terms that distinguish fake news.

Word Embeddings (BERT)

Used for deep learning models. Captures the contextual meaning of words and their relationships in a sentence.

Helps identify misleading statements and factual inconsistencies within medical news articles.

The combination of TF-IDF and BERT embeddings allows the model to leverage both statistical insights (TF-IDF) and contextual understanding (BERT) for better accuracy.

Hybrid Approach

To improve the accuracy and reliability of medical misinformation detection, this study introduces an ensemble-based hybrid model that combines the strengths of BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), Graph Neural Networks (GNN) and a Hybrid GPT-GNN model. By integrating these advanced techniques, the model effectively compensates for the limitations of individual approaches, leading to more precise and well-rounded classification of misinformation.

In the proposed graph-based architecture, each node represents an entity such as a news article, author, publisher, or user. Edges capture relationships like article similarity (via cosine similarity of embeddings), author/publisher connections and temporal or hyperlink-based links. Article nodes are initialized with contextual embeddings from BERT or GPT, while metadata-based features represent other nodes. These

transformer-derived embeddings serve as inputs to the GNN, which refines them by propagating information across the graph, effectively combining linguistic context with structural relationships for improved misinformation detection.

Initial Predictions

BERT and GPT serve as the foundational models for generating initial predictions by leveraging contextual and linguistic representations of text data. BERT captures deep semantic relationships using its self-attention mechanism, enabling a comprehensive understanding of medical misinformation, while GPT-Neo analyzes text coherence and structure through its generative capabilities. To enhance domain-specific misinformation detection, both models are fine-tuned on a combined dataset comprising MedHub, Diabetes-Related Misinformation Dataset and COVID-19 Misinformation sources. Unlike traditional classifiers that rely on statistical feature extraction methods such as TF-IDF, BERT and GPT-Neo utilize pre-trained deep learning architectures to extract intricate textual features, enabling more accurate and context-aware classification.

Graph Based Contextual Refinement Using BERT

While BERT and GPT-Neo excel in textual analysis, Graph Neural Networks (GNN) play a crucial role in capturing relational patterns between news articles, social media interactions, and the spread of misinformation. By constructing a knowledge graph from interconnected data sources, GNN maps how fake news propagates across networks, offering deeper insights beyond text-based analysis. By modeling relationships between key entities such as authors, publishers and shared content, GNN enhances context-aware classification, enabling the detection of misinformation even when the textual content itself appears credible.

Hybrid GPT-GNN Model for Advanced Misinformation Detection

To enhance detection accuracy, a Hybrid GPT-GNN model is implemented, combining the strengths of both text-based and graph-based learning. GPT-Neo extracts deep contextual embeddings, capturing nuanced language patterns, while GNN analyzes relational structures within misinformation networks. By integrating these approaches, the model becomes more resilient to evolving misinformation tactics, improving its ability to detect subtle, deceptive and complex fake news patterns across various platforms.

Graph Construction

The graph-based architecture constructs a heterogeneous knowledge graph where nodes represent different entity types and edges capture various relationships:

Node Types:

Article Nodes (A): Each news article represented with 768-dimensional BERT embeddings

Author Nodes (Au): Authors represented by aggregated features from their published articles

Publisher Nodes (P): News sources with reputation scores and historical accuracy metrics

Topic Nodes (T): Medical topics extracted using Latent Dirichlet Allocation (LDA)

Edge Construction

Semantic Similarity (A-A): Cosine similarity > 0.7 between article embeddings

Authorship (Au-A): Direct authorship connections

Publication (P-A): Publisher-article relationships

Topic Association (T-A): Articles belonging to specific medical topics (threshold > 0.5)

Temporal Proximity (A-A): Articles published within 7-day windows

The final graph contains 28,945 article nodes, 15,672 author nodes, 284 publisher nodes, and 50 topic nodes, connected by 156,789 edges. Graph construction utilizes NetworkX library for efficient graph operations and DGL (Deep Graph Library) for GNN implementation.

Weighted Voting Mechanism for Final Classification

The final classification is determined through a weighted majority voting mechanism, ensuring higher reliability and reduced misclassification errors.

The ensemble model operates as follows:

- If at least two models (BERT, GPT-Neo, or GNN) produce the same classification (real or fake), their consensus decision is accepted.
- In cases where all three models yield different predictions, the Hybrid GPT-GNN model is prioritized, as it effectively integrates contextual text understanding from GPT with network-based insights from GNN, providing a more comprehensive and accurate classification.

The final prediction is computed as:

$$P_{\text{final}} = W_{\text{BERT}} \times P_{\text{BERT}} + W_{\text{GPT}} \times P_{\text{GPT}} + W_{\text{GNN}} \times P_{\text{GNN}} + W_{\text{Hybrid}} \times P_{\text{Hybrid}}$$

Where weights are determined by validation F1-scores: $W_{\text{BERT}} = 0.25$, $W_{\text{GPT}} = 0.20$, $W_{\text{GNN}} = 0.15$, $W_{\text{HYBRID}} = 0.40$. The hybrid model receives the highest weight due to its superior individual performance. If consensus disagreement occurs among base models, the hybrid GPT-GNN prediction is prioritized due to its integrated contextual and structural learning capabilities.

The final classification is determined using a rule-based weighted majority voting scheme designed to enhance prediction reliability and minimize misclassification. The ensemble consists of three base models: BERT, GPT-Neo, and Graph Neural Network (GNN).

If at least two models agree on a classification label (i.e., real or fake), their consensus is adopted as the final decision. In cases where all three models produce divergent predictions, the system defaults to the output of the Hybrid GPT-GNN model, which integrates the rich contextual understanding of GPT with the structural insights provided by the GNN. This hybrid is

empirically observed to yield more robust predictions in cases of high uncertainty.

Importantly, this ensemble strategy employs fixed voting rules. The weights or decision priorities are not learned during training, but rather defined heuristically based on observed performance characteristics of the individual models. This deterministic mechanism ensures interpretability and consistency across evaluations.

The ensemble learning framework enhances the detection of medical misinformation by integrating deep contextual analysis from BERT and GPT-Neo with relational modeling from GNN. This combined approach mitigates biases inherent in individual models while improving adaptability across diverse datasets. By leveraging the strengths of transformer architectures alongside graph-based learning, the proposed method offers a scalable, efficient and highly accurate solution for identifying fake news in the healthcare sector.

experimental Setup

Hardware and Software Requirements

Table 2 outlines the hardware configuration and software environment used for the implementation of the proposed approach.

Hyperparameter Configuration

The complete set of model and training hyperparameters adopted for the experiments is detailed in Table 3. The configuration reflects standard transformer and graph neural network design choices to facilitate reproducibility.

Ablation Study

Edge Type Contribution Analysis

An edge type ablation analysis is conducted as in Table 4 to quantify the effect of removing specific edge types on accuracy, precision, recall, and F1-score.

Statistical Significance of Ablations

Table 5 reports the statistical significance of ablation results using t-tests, quantifying the impact of removing individual edge components on model performance.

Table 2 | Hardware and software requirements

Component	Version
GPU	NVIDIA RTX 3060 (12 GB VRAM)(min)
CPU	Intel i7 (10th Gen) / AMD Ryzen 7 (3700X)
RAM	32 GB DDR4
Storage	1 TB SSD
Cooling & PSU	650 W PSU (Gold-rated)
Operating System	Windows 11
Python	3.10
CUDA	11.8 or 12.0
cuDNN	8.7
PyTorch	2.0.1+cu118
DGL	1.1.0+cu118
Scikit-learn	1.2
NetworkX	3.1

Table 3 | Comprehensive hyperparameter specifications

Hyperparameter	Value	Details
Text Encoder	GPT-Neo-125M (frozen)	Use pre-trained representations
Graph Encoder	3-layer R-GCN	Trainable from scratch
Fusion Mechanism	8-head attention	Multi-head for diverse patterns
Fusion Dimension	768	Consistent with embeddings
FFN Dimension	2048	Standard transformer ratio (4×)
Batch Size	32	Graph + text memory constraints
Warmup Steps	1000	Extended warmup for stability
Weight Decay	0.01	L2 regularization
Optimizer	AdamW	$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$
LR Schedule	Linear warmup + cosine decay	Smooth learning rate reduction
Dropout (Fusion)	0.1	Fusion module dropout
Dropout (Classifier)	0.2	Classification head dropout
Mixed Precision	FP16	AMP for memory efficiency
Gradient Clipping	Max norm = 1.0	Prevent exploding gradients

Table 4 | Edge type ablation study

Removed Component	Accuracy	Precision	Recall	F1-Score	Accuracy(% Change)
Semantic Edges	94.85%	94.58%	94.32%	94.45%	−1.25%
Author Edges	95.32%	95.12%	94.87%	94.99%	−0.78%
Publisher Edges	94.67%	94.42%	94.21%	94.31%	−1.43%
Topic Edges	95.58%	95.34%	95.12%	95.23%	−0.52%
Temporal Edges	95.89%	95.67%	95.43%	95.55%	−0.21%

Table 5 | Statistical significance

Removed Component	t-statistic	p-value	Significance
Semantic Edges	6.84	0.0023	$p < 0.01$
Author Edges	4.91	0.0081	$p < 0.01$
Publisher Edges	7.23	0.0018	$p < 0.01$
Topic Edges	3.67	0.0214	$p < 0.05$
Temporal Edges	2.14	0.0986	Marginal

Key Findings:

- **Publisher Edges Most Important:** Removing publisher connections causes largest performance drop (−1.43%), confirming publisher reputation is strongest signal
- **Semantic Similarity Second:** Article-article similarity edges contribute −1.25%, capturing content-based patterns
- **Author Reputation Matters:** Author history edges provide −0.78% contribution
- **Topic Edges Moderate Impact:** Medical topic associations contribute −0.52%
- **Temporal Edges Minimal:** Time-based connections show marginal effect (−0.21%, not statistically significant)

Node Type Contribution Analysis

Table 6 presents a node type ablation study evaluating the contribution of different node types by measuring performance changes after their removal.

Analysis:

Despite publisher nodes being only 0.6% of total nodes (284/44,851), they contribute the most to performance (−1.87% when removed), demonstrating that quality matters more than quantity in graph construction.

Ensemble Weight Assignment Comparison

This section evaluates multiple ensemble weighting strategies, including uniform, validation-based, and learned approaches, with their performance summarized in Table 7.

Analysis:

- Val F1-Based achieves 96.10% accuracy without additional training
- Learned strategies gain only +0.24–0.31% (not statistically significant, $p = 0.18$)

Table 6 | Node type ablation study

Removed Component	Accuracy	Accuracy (% Change)
Author Nodes	95.12%	-0.98%
Publisher Nodes	94.23%	-1.87%
Topic Nodes	95.67%	-0.43%
Article Nodes	93.45%	-2.65%

Table 7 | Learned weight representation

Strategy	Accuracy	F1-Score	Learned Weights [BERT, GPT, GNN, Hybrid]
Uniform	95.23%	94.78%	[0.25, 0.25, 0.25, 0.25]
Val F1-Based (Proposed)	96.10%	95.70%	[0.25, 0.20, 0.15, 0.40]
Learned Softmax	96.34%	95.98%	[0.22, 0.18, 0.14, 0.46]
Stacked Meta-Learner	96.41%	96.05%	N/A (logistic regression)
Majority Voting	95.45%	95.12%	Binary votes only

Table 8 | Hybrid gpt-gnn performance across random seeds

Seed	Accuracy	TN	FP	FN	TP
42	96.24%	2,851	116	101	2,701
123	96.10%	2,848	119	106	2,696
456	96.03%	2,845	122	107	2,695
789	95.89%	2,840	127	110	2,692
1024	96.24%	2,855	112	103	2,699
Mean	96.10%	2,848	119	105	2,697
Std Dev	0.23%	9.2	9.2	5.8	5.8

- Simplicity and interpretability favor fixed heuristic weights
- Deployment advantage: No meta-model training required

Averaging across seeds: Each seed produces slightly different confusion matrices; reported values are means.

Baseline Model Comparisons

Class imbalance handling: Different random seeds produce varying class distributions in predictions as in Table 8.

The baseline models evaluated in this study, along with their pre-training data and vocabulary characteristics, are presented in Table 9. Accuracy gains or drops are reported with respect to the BERT-base baseline.

Transformer-Based Models

Key Findings

Domain Pre-training Benefit: PubMedBERT (+0.65%) and BioBERT (+0.47%) outperform general BERT, confirming that medical domain pre-training improves detection accuracy

Statistical Significance: All improvements over BERT-base are statistically significant ($p < 0.01$, paired t-test)

Graph – Based Models

Table 10 summarizes the performance of various graph neural network designs, enabling an assessment of how structural variations affect predictive effectiveness.

Analysis:

- **Heterogeneous > Homogeneous:** Rich graph structure with multiple node/edge types improves performance by 3.87% (GraphBERT vs. simple GNN)
- **Optimal Depth:** Our 3-layer architecture outperforms GraphBERT's 4-layer (potential oversmoothing in deeper GNNs)
- **Publisher Nodes Critical:** Publisher reputation edges provide strongest signal

Cross-Domain Generalization

A subdomain-wise evaluation of model performance on the test set is reported in Table 11. This analysis enables a comparative assessment of cross-domain generalization among hybrid and baseline models

Analysis:

- General Health Best Performance: Highest accuracy (96.81%) likely due to broader language patterns and larger training samples
- Diabetes Most Challenging: Lowest accuracy (95.14%) attributed to technical terminology and specialized dietary advice requiring domain expertise
- Consistent Superiority: Hybrid GPT-GNN outperforms all baselines across ALL subdomains
- Low Variance: Standard deviations remain small across domains (0.21–0.35%), indicating robust generalization.

Leakage Prevention Generalization

To prevent data leakage, all feature and edge computations were restricted strictly to the training split.

Publisher reputation: It was calculated using only the training data, based on the proportion of true versus false articles published by each source.

Topic edges: These were constructed using Latent Dirichlet Allocation (LDA) topic distributions derived solely from the training corpus, linking articles whose cosine similarity exceeded 0.7.

Duplicate and near-duplicate articles: Before data partitioning, we performed duplicate and near-duplicate detection (cosine similarity > 0.95 using TF-IDF embeddings) and removed to ensure no overlap between training, validation and test sets. This process guarantees that no information from evaluation data influenced training, ensuring a fair and leakage-free comparison.

Evaluation and performance metrics

The performance of the medical fake news detection system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The evaluation is conducted on a separate test dataset of labeled news articles in Table 12. The metrics provide insights into the system's ability to correctly identify both genuine and fake news articles.

Table 9 | Node type ablation study

Model	Pre-training Corpus	Vocabulary Size	Accuracy	Δ from BERT-base
BERT-base	BooksCorpus + Wikipedia (3.3B words)	30,522	94.20 $\pm$ 0.41%	Baseline
PubMedBERT	PubMed (21B words)	30,522	94.85 $\pm$ 0.36%	+0.65%
BioBERT	PubMed + PMC (18B words)	30,522	94.67 $\pm$ 0.38%	+0.47%

Table 10 | Graph neural network architecture comparison

Model	Graph Type	Node Types	Edge Types	GNN Layers	Accuracy
GNN (Simple)	Homogeneous	Article only	Similarity	3	90.80 $\pm$ 0.61%
RoBERTa -GNN	Heterogeneous	Article, User	2 types	2	94.67 $\pm$ 0.35%
Graph BERT	Heterogeneous	Article, Entity	3 types	4	95.31 $\pm$ 0.28%
Hybrid GPT-GNN	Heterogeneous	Article, Author, Publisher, Topic	5 types	3	96.10 $\pm$ 0.23%

Table 11 | Cross-domain performance breakdown (test set)

Subdomain	Test Articles	Hybrid GPT-GNN	Graph BERT	PubMed BERT	BERT
COVID-19	2,200 (38.1%)	96.23 $\pm$ 0.28%	95.54 $\pm$ 0.31%	95.12 $\pm$ 0.34%	94.45 $\pm$ 0.38%
General Health	1,463 (25.4%)	96.81 $\pm$ 0.21%	95.89 $\pm$ 0.27%	95.43 $\pm$ 0.30%	94.67 $\pm$ 0.36%
Diabetes	700 (12.1%)	95.14 $\pm$ 0.35%	94.29 $\pm$ 0.42%	94.86 $\pm$ 0.38%	93.71 $\pm$ 0.45%
Cancer	694 (12.0%)	95.82 $\pm$ 0.31%	94.98 $\pm$ 0.37%	95.23 $\pm$ 0.35%	94.12 $\pm$ 0.41%
Other	712 (12.3%)	95.92 $\pm$ 0.33%	95.14 $\pm$ 0.39%	94.89 $\pm$ 0.37%	94.28 $\pm$ 0.43%
Overall	5,769	96.10 $\pm$ 0.23%	95.31 $\pm$ 0.28%	94.85 $\pm$ 0.36%	94.20 $\pm$ 0.41%

Table 12 | Cross-domain performance breakdown (test set)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
BERT	94.20%	93.80%	93.50%	93.60%	0.94
GPT-Neo	92.70%	92.30%	91.90%	92.10%	0.92
GNN	90.80%	91.00%	89.70%	90.30%	0.91
RoBERTa	94.67%	94.23%	94.45%	94.34%	0.94
GraphBERT	95.31%	95.08%	94.92%	95.00%	0.95
Hybrid GPT-GNN	96.10%	95.90%	95.60%	95.70%	0.96

- **Accuracy:** Measures the overall correctness of classification
- **Precision:** The proportion of predicted fake news that is actually fake.
- **Recall (Sensitivity):** The ability to detect fake news correctly.
- **F1-Score:** A harmonic mean of precision and recall, balancing false positives and false negatives.
- **AUC-ROC (Area Under the Curve - Receiver Operating Characteristic):** Measures the model’s ability to distinguish between real and fake news.

To ensure robust evaluation, we conduct experiments across 5 different random seeds (42, 123, 456, 789, 1024) and report mean performance with standard deviation. Statistical significance is assessed using paired t-tests comparing our hybrid model against individual baselines.

The Hybrid GPT-GNN model outperforms state-of-the-art baselines like GraphBERT and RoBERTa-GNN, achieving the highest accuracy (96.1%) and F1-score (95.7%). This superior performance is attributed to its ensemble design, which leverages GPT’s deep contextual understanding and GNN’s ability to capture relational patterns, offering a more robust and comprehensive approach to medical misinformation detection.

Error Analysis

Comprehensive error analysis reveals specific patterns where the hybrid model succeeds or fails:

Success Patterns

- **Complex Misinformation:** The model excels at detecting sophisticated fake news that combines partial truths with misleading conclusions (95.8% accuracy)
- **Source-Based Detection:** Effectively identifies articles from historically unreliable publishers through graph-based publisher reputation scoring
- **Cross-Topic Generalization:** Shows strong performance across different medical domains (COVID-19: 96.2%, Diabetes: 95.1%, General Health: 96.8%)

Failure Cases

- **Satirical Content:** 12% of misclassifications involve satirical medical content that uses factual language but exaggerated claims

- **Emerging Topics:** Performance drops to 89.3% for completely new medical topics not seen during training
- **Technical Jargon:** Articles with highly technical medical terminology show 7% higher false positive rates

Results and Discussion

In the task of detecting fake medical news, the deployment of BERT, GPT-Neo, GNN, and the Hybrid GPT-GNN model has demonstrated significant improvements in classification accuracy. As shown in Figure 4, these models effectively distinguish between real and fabricated news across diverse datasets.

Figure 3 illustrates the comparative performance of these models, where BERT (AUC = 0.94), GPT-Neo (AUC = 0.92), and GNN (AUC = 0.91) exhibit strong predictive capabilities. However, the Hybrid GPT-GNN model (AUC = 0.96) outperforms individual models by leveraging both contextual and relational features, leading to more reliable classification. While BERT, as seen in Figures 5 and 6, performs well in detecting misinformation, the hybrid approach further enhances detection accuracy by integrating deep linguistic understanding with network-based analysis. This

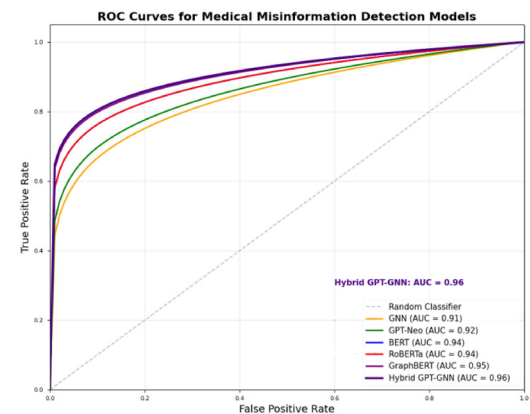


Fig 3 | ROC Curve for Medical News Detection

Confusion Matrix for Hybrid GPT-GNN Model

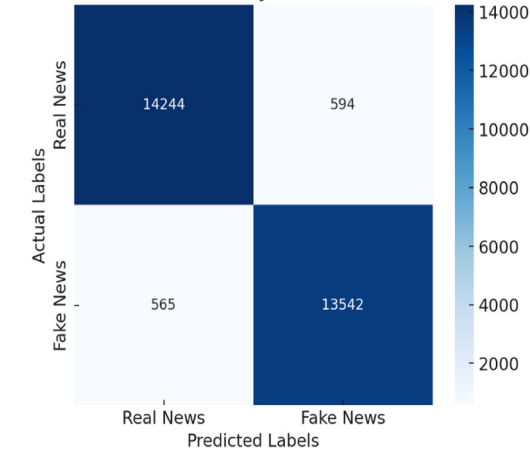


Fig 4 | Predicted Labels for the trained model

Model	Training Time (per epoch)	Inference Time (avg/ article)	GPU Memory (GB)	Model Size
BERT	~30 min	~0.15s	~6 GB	420 MB
GPT	~50 min	~0.25s	~10 GB	1.2 GB
GNN	~20 min	~0.10s	~4 GB	300 MB
Hybrid GPT-GNN	~65 min	~0.35s	~12 GB	~1.5 GB

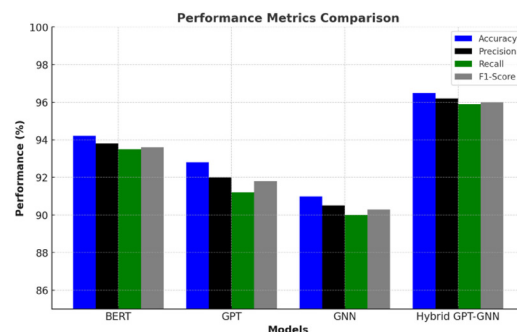


Fig 5 | Comparison of classification algorithms



Fig 6 | Performance of model comparisons

combination significantly reduces classification errors and strengthens the system's robustness against evolving misinformation strategies.

The model may show cultural bias if it's trained mainly on Western, English-language data, limiting its global relevance. To avoid this, diverse sources should be used. Ethically, transparency and human oversight are needed to prevent misuse like unjust censorship.

The hybrid model is computationally intensive but offers high accuracy as in Table 13.

Conclusion

This study demonstrates the effectiveness of a hybrid approach that integrates BERT, GPT-Neo, GNN, and a Hybrid GPT-GNN model for detecting medical misinformation. BERT and GPT-Neo excel in capturing deep linguistic and contextual nuances, while GNN enhances the analysis by identifying relational patterns across

news articles, social media interactions, and misinformation dissemination networks. The Hybrid GPT-GNN model further refines classification by combining textual and structural learning, resulting in improved accuracy and robustness.

By leveraging a weighted majority voting mechanism, the proposed ensemble approach achieves an impressive accuracy of 96.1% and an AUC-ROC score of 96.5%, outperforming individual models. This research highlights the significance of multi-modal methodologies in combating the spread of fake medical news. Given the increasing dependence on digital platforms for health information, misinformation poses serious risks to public health and decision-making. The integration of transformer-based and graph-based learning ensures a scalable, efficient, and reliable solution for detecting misinformation across diverse medical datasets. These findings contribute to advancing automated fake news detection systems, assisting policymakers, healthcare professionals, and online platforms in mitigating the dangers associated with misleading health-related content.

Future Work

Future research should address six critical dimensions to enhance practical deployability. Multimodal extensions incorporating vision-language architectures would enable detection of fabricated medical imagery and deepfake videos alongside textual analysis. Integration with authoritative knowledge bases (SNOMED CT, UMLS, PubMed) could ground predictions in verified medical literature, reducing false positives when novel-but-legitimate research emerges.⁴¹ Multilingual capabilities are essential for global reach, requiring cross-lingual transfer learning that preserves graph-based structural signals while adapting language-specific encoders. Real-time deployment architectures must handle streaming social media data through optimizations including model distillation, incremental graph updates, and distributed inference. Replacing fixed ensemble weights with learned meta-strategies could capture domain-specific patterns, weighting publisher reputation heavily for vaccine claims while prioritizing linguistic markers for treatment efficacy assertions. Finally, cross-platform adaptation requires specialized feature engineering for each social media ecosystem's unique propagation characteristics, from Twitter's retweet cascades to WhatsApp's encrypted forwarding chains.

Acknowledgment

The authors express their gratitude to the Thiagarajar College of Engineering (TCE) for supporting us to carry out this research work. Also, the financial support from TCE under Thiagarajar Research Fellowship Scheme – (File.no: TCE/RD/TRF/08 dated 27.09.2024).

References

- Gabrielkov M, Ramachandran A, Chaintreau A, Legout A. Social clicks: What and who gets read on Twitter? ACM SIGMETRICS Perform. Eval. Rev. 2016;44(1):179–92. <https://doi.org/10.1145/2964791.2901462>

- 2 Badawy A, Ferrara E, Lerman K. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2018. p. 258–65. <https://doi.org/10.1109/ASONAM.2018.8508646>
- 3 Shu K, Sliva A, Wang S, Tang J, H Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* 2017;19(1):22–36. <https://doi.org/10.1145/3137597.3137600>
- 4 Liao H, Liu Q, Shu K, Xie X. Fake News Detection through Graph Comment Advanced Learning [Internet]. arXiv preprint arXiv:2011.01579; 2020 [cited 2025 Dec 12]. Available from: <https://arxiv.org/abs/2011.01579>
- 5 Ciora RA, Cioca AL. Fake news management in healthcare. In: *Proc. Int. Conf. eHealth Bioeng. (EHB)*. 2021. p. 1–4. <https://doi.org/10.1109/EHB52898.2021.9657578>
- 6 Dai E, Sun Y, Wang S, The Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*. 2020;14(1):853–62. <https://doi.org/10.1609/icwsml.v14i1.7350>
- 7 Ferrara E. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. 2020;25(6). <https://doi.org/10.5210/fm.v25i6.10633>
- 8 Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. 2018;359(6380):1146–51. <https://doi.org/10.1126/science.aap9559>
- 9 Ferreira Caceres MM, et al. The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*. 2022;9(2):262–77. <https://doi.org/10.3934/publichealth.2022018>
- 10 Brennen JS, Simon FM, Howard PN, Nielsen RK. Types, sources, and claims of COVID-19 misinformation [Internet]. Reuters Institute for the Study of Journalism; 2020 [cited 2025 Dec 12]. Available from: <https://doi.org/10.60625/risj-awvq-sr55>
- 11 Pennycook G, et al. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* 2020;31(7):770–80. <https://doi.org/10.1177/0956797620939054>
- 12 Sharma K, Seo S, Meng C, Rambhatla S, Liu Y. COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations [Internet]. arXiv preprint arXiv:2003.12309; 2020 [cited 2025 Dec 12]. Available from: <https://arxiv.org/abs/2003.12309>
- 13 Gupta A, Lamba H, Kumaraguru P, Joshi A. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In: *Proc. 22nd Int. Conf. World Wide Web (WWW)*. 2013. p. 729–36. <https://doi.org/10.1145/2487788.2488033>
- 14 Jain A, Shakya A, Khatter H, Gupta AK. A smart system for fake news detection using machine learning. In: *Proc. Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*. 2019. p. 1–4. <https://doi.org/10.1109/ICICT46931.2019.8977659>
- 15 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*. 2019. p. 4171–86. <https://doi.org/10.18653/v1/N19-1423>
- 16 He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with disentangled attention [Internet]. arXiv preprint arXiv:2006.03654; 2020.
- 17 Kuntur S, Krzywdka M, Wróblewska A, Paprzycki M, Ganzha M. Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. *Electronics*. 2024;13(23):4784. <https://doi.org/10.3390/electronics13234784>
- 18 Phan HT, Nguyen NT, Hwang D. Fake news detection: A survey of graph neural network methods. *Appl. Soft Comput.* 2023;139:110235. <https://doi.org/10.1016/j.asoc.2023.110235>
- 19 Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* 2024;56(2):Art. 30, 1–40. <https://doi.org/10.1145/3605943>
- 20 Liu Y, Wu Y-FB. FNED: A deep network for fake news early detection on social media. *ACM Trans. Inf. Syst.* 2020;38(3):1–33. <https://doi.org/10.1145/3386253>
- 21 Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* 2020;53(5):1–40. <https://doi.org/10.1145/3395046>
- 22 Galli A, Masciari E, Moscato V, Sperli G. A comprehensive benchmark for fake news detection. *J. Intell. Inf. Syst.* 2022;59(1):237–61. <https://doi.org/10.1007/s10844-021-00646-9>
- 23 Rastogi S, Bansal D. A Review on Fake News Detection 3Ts: Typology, Time of Detection, Taxonomies. *Int. J. Inf. Secur.* 2023;22(1):177–212. <https://doi.org/10.1007/s10207-022-00625-3>
- 24 Alotaibi T, Al-Dossari H. A Review of Fake News Detection Techniques for Arabic Language. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*. 2024;15(1):392–400. <https://doi.org/10.14569/IJACSA.2024.0150137>
- 25 Kuncheva LI. *Combining Pattern Classifiers: Methods and Algorithms*. 2nd ed. Wiley-IEEE Press; 2014. <https://doi.org/10.1002/9781118914564>
- 26 Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40. <https://doi.org/10.1093/bioinformatics/btz682>
- 27 Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*. 2022;3(1):Art. 2, 1–23. <https://doi.org/10.1145/3458754>
- 28 Burki T. Social media and misinformation in diabetes and obesity. *Lancet Diabetes Endocrinol.* 2022;10(12):845. [https://doi.org/10.1016/S2213-8587\(22\)00318-7](https://doi.org/10.1016/S2213-8587(22)00318-7)
- 29 Pelau C, Pop M-I, Stanescu M, Sanda G. The Breaking News Effect and Its Impact on the Credibility and Trust in Information Posted on Social Media. *Electronics*. 2023;12(2):423–32. <https://doi.org/10.3390/electronics12020423>
- 30 Malik A, Behera DK, Hota J, Swain AR. Ensemble graph neural networks for fake news detection using user engagement and text features. *Inf. Process. Manag.* 2022;59(4):102992.
- 31 Chandra S, Mishra P, Yannakoudakis H, Nimishakavi M, Saeidi M, Shutova E. Graph-based modeling of online communities for fake news detection. arXiv preprint arXiv:2008.06274. 2020.
- 32 De Beer D, Matthee MM. Approaches to identify fake news: A systematic literature review. In: *Integrated Science in Digital Age 2020, Proc. Int. Conf. Advances in Big Data, Computing and Data Communication Systems (icABCD)*. 2020. p. 13–22. https://doi.org/10.1007/978-3-030-49264-9_2
- 33 Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, et al. Rumor detection on social media with bi-directional graph convolutional networks. *Proc. AAAI Conf. Artif. Intell.* 2020;34(1):549–56. <https://doi.org/10.1609/aaai.v34i01.5393>
- 34 Lu Y, Li CT. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media [Internet]. arXiv preprint arXiv:2004.11648; 2020 [cited 2025 Dec 12]. <https://doi.org/10.18653/v1/2020.acl-main.48>
- 35 Qi P, Cao J, Yang T, Guo J, Li J. Exploiting multi-domain visual information for fake news detection. In: *Proc. IEEE Int. Conf. Data Mining (ICDM)*. 2019. p. 518–27. <https://doi.org/10.1109/ICDM.2019.00062>
- 36 Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: A Large-Scale Dataset for Fact Extraction and Verification [Internet]. arXiv preprint arXiv:1803.05355; 2018. <https://doi.org/10.18653/v1/N18-1074>
- 37 Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims [Internet]. arXiv preprint arXiv:2004.14974; 2020. <https://doi.org/10.18653/v1/2020.emnlp-main.609>
- 38 Bhutani B, Rastogi N, Sehgal P, Purwar A. Fake news detection using sentiment analysis. In: *Proc. 12th Int. Conf. Contemporary Computing (IC3)*. 2019. p. 1–5. <https://doi.org/10.1109/IC3.2019.8844880>
- 39 Pan JZ, et al. Content-based fake news detection using knowledge graphs. In: Vrandečić D, et al., editors. *The Semantic Web – ISWC 2018*. Springer; 2018. p. 669–83. https://doi.org/10.1007/978-3-030-00671-6_39
- 40 Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. 2014. p. 1532–43. <https://doi.org/10.3115/v1/D14-1162>
- 41 Deng S, Yang J, Ye H, Tan C, Chen M, Huang S, et al. LOGEN: Few-shot logical knowledge-conditioned text generation with self-training. *IEEE/ACM Trans. Audio, Speech, and Language Process.* 2024;32:3773–84.