Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Contextual Stance-Aware Semantic Graph Learning for Fake News Detection

Djamila Benchikh [†], Mohand Saïd Allili [†,*], Etienne Gael Tajeuna [†]
[†] University of Quebec in Outaouais, Gatineau, Quebec, Canada.

Abstract

The rapid spread of disinformation on social networks threatens public trust and democratic processes. We propose a unified framework for early detection of emerging false narratives by combining context-aware graph modeling with semantic analysis. Our method builds interaction graphs where posts are nodes connected by stance relations (agreement or disagreement). In parallel, a semantic module extracts fine-grained linguistic cues from each post. These signals are fused via a graph neural network that jointly models early diffusion patterns and content semantics to identify deceptive posts at their inception. Experiments on benchmark datasets show that our approach outperforms existing baselines, highlighting the effectiveness of integrating stance-aware graph representations with semantic understanding for scalable disinformation detection.

Keywords: Fake news detection, Graph neural networks (GNNs), Stance-aware modeling, Semantic discourse analysis.

1. Introduction

Disinformation and fake news pose an escalating threat across social media platforms, undermining public trust and safety [1]. Empirical studies show that false information spreads faster and reaches more people than truthful content [2]. Diffusion analyses attribute this to factors such as novelty, emotional arousal and homophily [3]. Although fake-news consumption is concentrated among a small subset of users, its political impact and targeted distribution give it disproportionate influence [4]. Coordinated inauthentic behavior and social bots further amplify low-credibility content at early stages, before fact-checking can intervene [5]. The public-health risks became evident during COVID-19, when the World Health Organization described an "infodemic," with global studies documenting the scale and harm of pandemic-related misinformation [6].

Existing disinformation detection methods include content-based, propagation-based, and hybrid approaches. Content-based models rely on linguistic cues and miss cross-post dynamics such as agreement or contradiction [1]. Social-context methods model interactions through graphs but depend on explicit links that are often sparse or incomplete [7]. Hybrid models [8] treat relations as homogeneous, overlooking the argumentative structure of misinformation. However, fake news spreads not only through content but also through latent relational structures among posts, often implicit via semantic similarity or stance alignment rather than explicit interactions [2]. This motivates modeling posts as a graph where edges are induced from semantic similarity and enriched with stance-aware signals [9]. Such graphs capture latent communities of reinforcing or conflicting posts, as well as agreement–disagreement dynamics, while integrating temporal, linguistic, and user-level context. More broadly, graph-based models capture complex dependencies and propagation patterns and support temporal dynamics for early detection [7].

In this paper, we introduce a unified stance-aware framework where the interaction graph is implicitly constructed from semantic similarity between posts. A transformer-based module produces contextual embeddings, while a stance detection component assigns agreement or disagreement weights to the induced edges between semantically close posts. The resulting

[*] mohandsaid.allili@uqo.ca

graph is processed by a GNN that jointly leverages textual semantics, latent relations, and argumentative cues for early detection of deceptive narratives. The main contributions are as follows: (1) We propose a stance-aware implicit graph construction method, where edges are induced by both semantic similarity and argumentative interactions, capturing agreement and disagreement patterns often ignored by existing approaches. (2) We enrich node representations by integrating semantic and user-level features with explicit stance information, enabling a more comprehensive modeling of credibility and discourse dynamics. (3) Extensive experiments on the TruthSeeker [10] and PHEME [11] datasets, with comparison to state-of-the-art methods, show that our approach significantly outperforms content-only and structure-only baselines, demonstrating the effectiveness of using stance-based signals for more effective disinformation detection.

The rest of the paper is organized as follows: Section 3 details the proposed framework. Section 4 presents the experimental setup, dataset, and evaluation results. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related work

Online *disinformation* aims to manipulate public opinion, often through coordinated campaigns and strategic amplification. Empirical studies show that false narratives propagate farther, faster, and more broadly than truthful information on social platforms. Consequently, modern approaches model disinformation within its social context by integrating user behavior and temporal diffusion patterns. Fake news detection is commonly formulated as claim or document veracity classification [2], with representative benchmarks such as LIAR and FakeNewsNet [1]. Rumor detection instead focuses on evolving discussion threads and early signals [12].

Graph-based modeling has become central to disinformation detection, as social media naturally forms relational structures linking users, posts, and propagation paths [13]. Propagation-based models such as Bi-directional GCN learn complementary top-down and bottom-up diffusion patterns [9], while user-aware frameworks (e.g., UPFD [14]) show that interaction preferences enhance detection when combined with textual semantics. Recent advances leverage heterogeneous and attention-based GNNs to fuse multi-relational evidence [8], but challenges remain in early detection under sparse diffusion and in mitigating spurious social correlations [7]. Hybrid approaches address this by combining semantic encoding with graph reasoning and uncertainty modeling [15]. Stance information further bridges textual semantics and social dynamics and has become a key signal in rumor verification. This motivates *stance-aware semantic graphs* that propagate stance consistency while aligning structural and semantic evidence. Our approach explicitly models these dynamics by capturing agreement and contradiction relationships between posts within a unified graph framework, thereby enabling early detection even before large-scale propagation occurs.

3. Methodology

Our methodology represents message dissemination as a heterogeneous graph where posts are linked not only by interactions but also by agreement and contradiction, enabling joint reasoning over narrative alignment and textual meaning. Unlike approaches that decouple text and structure, our method performs unified semantic–structural inference, supporting early detection before large cascades form. The overall model is shown in Fig. 1.

3.1. Graph construction and stance representation

The foundation of our approach lies in the construction of a post interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v \in \mathcal{V}$ represents an individual post. Each edge $e_{ij} \in \mathcal{E}$ captures an

interaction between posts p_i and p_j, such as that the posts carry the same semantic content, typically occurring in reply, retweet, quote, or mention actions.

Each post is first preprocessed by removing duplicates and non-informative elements, then represented as a node in a graph. Edges are established based on semantic similarity or interactions: an edge is created if the similarity exceeds a threshold T (typically $T = 0.8$, selected via grid search). This structure models information as a dynamic network capturing both social and semantic relations around a claim. To enrich the graph, we extract contextual embeddings for each post using SBERT [16] and use them as node features alongside metadata (e.g., user credibility, number of followers). Edges are then defined using cosine similarity between embeddings, with the threshold controlling graph density.

To model interactions, each edge carries a stance label (*agreement* vs. *disagreement*) and a weight w_{ij} ($w_{ij} = 1$ for agreement, $w_{ij} = -1$ for disagreement), yielding a graph that jointly encodes diffusion and contention. This departs from prior propagation models that treat links as homogeneous. Unlike methods that aggregate stance at the user or path level, we assign stance directly to edges for finer-grained modeling. This enables the capture of structural patterns such as polarization and cross-community bridges, known to be discriminative for misinformation, and provides a richer relational bias than attention- or position-based weighting alone [15]. To model semantic agreement and contradiction between posts, we use RoBERTa [17]. The task is formulated as binary sentence classification, where each input pair (p_i, p_j) is labeled as agreement or disagreement. The posts are concatenated following the sequence format:

$$\texttt{[CLS]}\ p_i\ \texttt{[SEP]}\ p_j\ \texttt{[SEP]},$$

where `[CLS]` is the classification token and `[SEP]` separates posts. The joint representation is taken from the final hidden state of `[CLS]` and passed to a linear layer with cross-entropy loss to predict agreement or contradiction. Fig. 2 shows graphs from the TruthSeeker dataset [10], highlighting structural differences between misinformation and authentic networks. Fake news graphs exhibit bursts of disagreement, often driven by low-credibility users, whereas authentic news shows more agreement across users with varying credibility.

3.2. **Graph-Level representation and classification**

After contextual encoding, each graph $G = (V, E)$ consists of nodes $v_i \in V$ representing posts and edges $e_{ij} \in E$ denoting semantic or argumentative relations. Each node v_i is initialized with a feature vector $x_i \in \mathbb{R}^d$, obtained by concatenating (i) the semantic embedding of the tweet text from SBERT and (ii) user-level metadata such as credibility, bot-likelihood, or influence. Edges e_{ij} are annotated with attributes a_{ij} indicating stance (agreement vs. disagreement). These attributed graphs are processed by a Graph Neural Network using message passing. At each layer l, the representation of node v_i is updated by aggregating information from its neighbors $N(i)$:

$$h_i^{(l)} = \sigma\Big(W^{(l)} \cdot \mathrm{AGG}^{(l)}\big\{h_j^{(l-1)}, a_{ij} \mid j \in N(i)\big\}\Big), \tag{3.1}$$

where AGG is an aggregation function (e.g., mean or attention), $W^{(l)}$ is a trainable weight matrix, a_{ij} is the edge attribute (agreement/disagreement), and σ is an activation (e.g., ReLU). Through L layers, nodes accumulate semantic, social, and argumentative context from their L-hop neighborhood.

A global readout operation (e.g., mean attention pooling) aggregates the final node embeddings $\{h_i^{(L)}\}$ into a graph-level representation $z_G \in \mathbb{R}^k$, where $z_G = \mathrm{READOUT}\big(\{h_i^{(L)} \mid i \in V\}\big)$, as a Global pooling layer. This graph encodes (i) semantic content, (ii) argumentative dynamics, and (iii) user credibility or influence signals. Finally, z_G is passed to a MLP discriminator to predict veracity $\hat{y}_G = \mathrm{MLP}(z_G)$. The full network is trained end-to-end using the binary cross-entropy (BCE) loss function.

4. Experimental results

4.1. Datasets

Our experiments used mainly TruthSeeker dataset [10], which is one of the largest corpora for studying misinformation. It contains tweets related to both true and false news events, each linked to fact-checked press event from the PolitiFact corpus [18]. The dataset includes user posts expressing opinions, shares, or reactions to the news items. For *True content* (class 1), it has 68,915 posts around 479 news, whereas for *False content* (class 0) it has 65,267 posts around 479 news. Note that the tweet density per news item is, on average, a higher for false news (136.25) than for true news (119.02). This reflects a trend already noted in the literature where misleading content tends to attract greater engagement and spread more rapidly across social networks. To assess the robustness of our framework beyond our primary dataset, we conducted a complementary evaluation on the popular PHEME benchmark [11], a widely used dataset for event-driven rumour detection and related veracity tasks. Results of the the PHEME dataset are shown in Appendix C. Note that for each dataset, we have split the data into two subsets: 80% for training and 20% for test. Results are reported for the test part only.

4.2. Experimental protocol

Our evaluation is mainly based on the TruthSeeker dataset, where we progressively integrate different components of the proposed framework to assess their individual and combined contributions. We compare the following configurations: 1) *Baseline (Text-only)* [19]: where each post is encoded with SBERT embeddings, 2) *Graph-only (SBERT nodes)* [20]: where posts are represented as nodes with SBERT embeddings, and graphs are constructed implicitly via semantic similarity. 3) *Node feature enrichment* [10]: where user attributes are added to node representations. 4) *News-level embedding*: using semantic embedding of the associated news to provide global contextual information. 5) *Full stance-aware graph model*: where edge encoding (agreement vs. disagreement) is integrated into the graph.

Experiment #1: Node semantics were encoded using SBERT-large embeddings (768 dimensions), without explicit tweet–tweet relations or edge attributes. This setup isolates the contribution of textual embeddings within a graph framework. Two approaches were compared: (1) SBERT aggregated using GCN, GraphSAGE, or GAT for veracity prediction; (2) a text-only baseline where embeddings were averaged per news item and classified with an MLP. All GNN models outperformed the baseline (61.32%). GCN achieved the highest accuracy (81.13%), followed by GraphSAGE (80.45%) and GAT (79.80%). These results show that even without edge attributes, graph structure improves the capture of interaction patterns, while text-only aggregation fails to model these relational dynamics.

Table 1. Performance comparison for Experiment #2.

Model configuration	**Accuracy**
SBERT representations only (tweets)	81.13%
SBERT + temporal features	83.96%
SBERT + temporal + linguistic features	88.68%
SBERT + temporal + linguistic + user attributes	89.62%
SBERT + all above + news-level embedding	**90.57%**

Experiment #2: Graph nodes were enriched with additional features, including linguistic attributes (e.g., number of adjectives, verbs) and user metadata (e.g., credibility, timestamp). These were combined with SBERT embeddings of the associated news item. Table 1 shows that accuracy improves as temporal, linguistic, and user features are progressively added. The best performance (90.57%) is obtained when all node-level features are

combined with semantic embeddings, confirming that rich contextual information enhances graph-based misinformation detection.

Experiment #3: We extend the most comprehensive node configuration by combining SBERT tweet embeddings, temporal metadata, linguistic features, user attributes, and news-level semantics, and further enrich the graph with stance-aware edges. As shown in Table 2, this significantly improves performance, reaching **93.87%** accuracy, the highest in our experiments. These results show that enriching node features beyond text substantially boosts performance (90.57% with news embeddings), while incorporating agreement and disagreement relations further enhances accuracy. This confirms that stance-aware relational signals are key to capturing discourse dynamics for effective misinformation detection.

Table 2. Performance with and without edge stance encoding: Experiment #3.

Model configuration	**Accuracy**	**F1-score**
SBERT + temporal + linguistic + user features + news embedding (no edge encoding)	90.57%	90.58%
Same configuration with argumentative edge encoding	**93.87%**	**93.86%**

Experiment #4: To contextualize the gains of our graph-based framework, we performed a cross-domain evaluation by fine-tuning a BERT-base classifier on PolitiFact and testing on TruthSeeker. As shown in Table 3, this setup achieves 75.47% accuracy and 71.11% F1, serving as a strong baseline. However, a large gap remains compared to our approach. While BERT relies on local textual cues, our model captures structured interactions and discourse-level dependencies. Our best configuration (Experiment #3), combining SBERT features with temporal, linguistic, and user attributes and stance-aware edges, achieves 93.87% accuracy and 93.86% F1. The +18.40 improvement highlights the importance of relational and stance-aware signals, demonstrating the advantage of modeling agreement and disagreement over text-only transfer methods.

Table 3. Text-only baselines: in-domain validation vs. cross-domain transfer.

Setting	**Accuracy**	**F1**
BERT-base (in-domain, filtered dataset → val)	75.34%	59.63%
BERT-base (cross-domain, PolitiFact → TruthSeeker test)	75.47%	71.11%

5. Conclusions

We presented a graph-based framework for fake news detection that models semantic, contextual, and argumentative aspects of posts. Experiments show that enriching nodes with linguistic, temporal, and user features significantly improves performance, while encoding inter-post stance relations further boosts accuracy. These findings highlight the importance of jointly capturing node semantics and edge-level discourse dynamics. The proposed approach provides a scalable and robust alternative to purely textual methods, supporting structure-aware learning. Future work will explore temporal evolution, cross-event generalization, and the integration of multimodal data (e.g., images/videos [21] and audio [22]), as well as user behavior and community dynamics to better understand misinformation propagation.

Acknowledgements

This work is supported by the Natural Sciences and Eng. Research Council of Canada.

References

[1] X. Zhou and R. Zafarani. "A survey of fake news: Fundamental theories, detection methods, and opportunities". In: *ACM Computing Surveys* 53.5 (2020), pp. 1–40.

[2] N. Capuano et al. "Content-Based Fake News Detection With Machine and Deep Learning: A Systematic Review". In: *Neurocomputing* 530 (2023), pp. 91–103.

[3] W. Ansar and S. Goswami. "Combating the menace: A survey on characterization and detection of fake news from a data science perspective". In: *Int'l J. of Information Management Data Insights* 1.2 (2021), p. 100052.

[4] D. Antypas, A. Preece, and J. Camacho-Collados. "A Multi-Faceted NLP Analysis of Misinformation Spreaders in Twitter". In: *14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. 2024, pp. 71–83.

[5] J. Pfander and S. Altay. "Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements". In: *Nature Human Behaviour* 9 (2025), pp. 688–699.

[6] S. Bhattacharya and A. Singh. "Unravelling the infodemic: a systematic review of misinformation dynamics during the COVID-19 pandemic". In: *Fron. in Communications* 10 (2025).

[7] G. Kątek et al. "In depth analysis for securing the truth: Addressing the fake news challenge with graph neural networks". In: *Neurocomputing* 654 (2025), p. 131327.

[8] C.-O. Truică, E.-S. Apostol, M. Marogel, and A. Paschke. "GETAE: Graph Information Enhanced Deep Neural Network Ensemble Architecture for fake news detection". In: *Expert Systems with Applications* 275 (2025), p. 126984.

[9] T. Bian, et al. "Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks". In: *AAAI*. 2020, pp. 549–556.

[10] S. Dadkhah et al. "The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker". In: *IEEE Trans. on Computational Social Systems* 11.3 (2024), pp. 3376–3390.

[11] E. Kochkina et al. *PHEME dataset for Rumour Detection and Veracity Classification*. 2018.

[12] S. Wu, Y. Deng, J. Liu, X. Luo, and G. Sun. "Rumor detection on social networks based on Temporal Tree Transformer". In: *PLOS ONE* 20.4 (2025), e0320333.

[13] H. T. Phan, N. T. Nguyen, and D. Hwan. "Fake news detection: A survey of graph neural network methods". In: *Applied Soft Computing* 139 (2023), p. 110235.

[14] X. Su et al. "Hy-DeFake: Hypergraph neural networks for detecting fake news in online social networks". In: *Neural Networks* 187 (2025), p. 107302.

[15] M. Ma, C. Zhang, Y. Li, J. Chen, and X. Wang. "Rumor detection model with weighted GraphSAGE focusing on node location". In: *Scientific Reports* 14 (2024), p. 27127.

[16] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Conference on Empirical Methods in Natural Language*. 2019, pp. 3982–3992.

[17] Y. Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. CoRR, abs /1907.11692. 2019.

[18] W. Y. Wang. "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection". In: *Annual Meeting of the Association for Computational Linguistics*. 2017, pp. 422–426.

[19] K. Yu, S. Jiao, and Z. Ma. "Fake News Detection Based on BERT Multi-domain and Multimodal Fusion Network". In: *Computer Vision and Image Understanding* 252 (2025), p. 104301.

[20] H. R. Moorthy, N. J. Avinash, N. S. K. Rao, K. R. Raghunandan, R. Dodmane, J. J. Blum, and L. A. Gabralla. "Dual Stream Graph Augmented Transformer Model Integrating BERT and GNNs for Context-Aware Fake News Detection". In: *Scientific Reports* 15 (2025), p. 25436.

[21] S. Belguesmia, M. S. Allili, and A. Hamadene. "Unmasking Facial DeepFakes: A Robust Multiview Detection Framework for Natural Images". In: *arXiv preprint arXiv:2510.15576* (2025). DOI: 10.48550/arXiv.2510.15576.

[22] D. E. Temmar et al. "Phonetic Analysis of Real and Synthetic Speech Using HuBERT Embeddings: Perspectives for Deepfake Detection". In: *IEEE SMC*. 2025, pp. 86–91.

[23] J. Dougrez-Lewis, E. Kochkina, M. Liakata, and Y. He. "Knowledge Graphs for Real-World Rumour Verification". In: *Proceedings of LREC-COLING 2024*. CC BY-NC 4.0. Torino, Italy: ELRA Language Resource Association, May 2024, pp. 9843–9853.

[24] P. Farinneya, M. M. Abdollah Pour, S. Hamidian, and M. Diab. "Active Learning for Rumor Identification on Social Media". In: *EMNLP (Findings)*. 2021, pp. 4556–4565.

Appendix A. Figure representing the overall proposed method

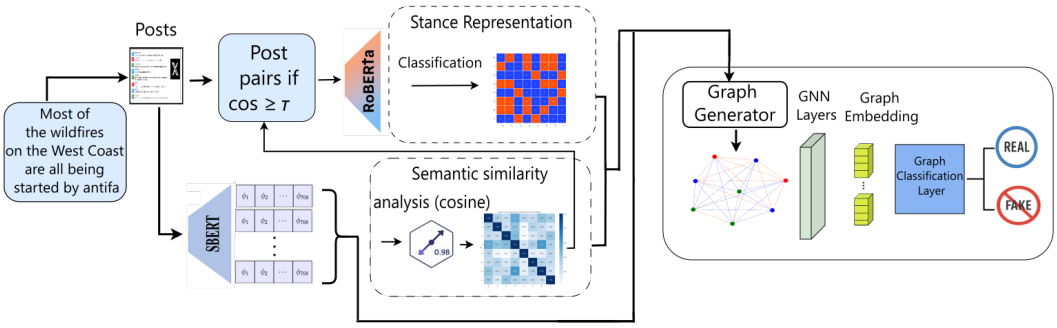
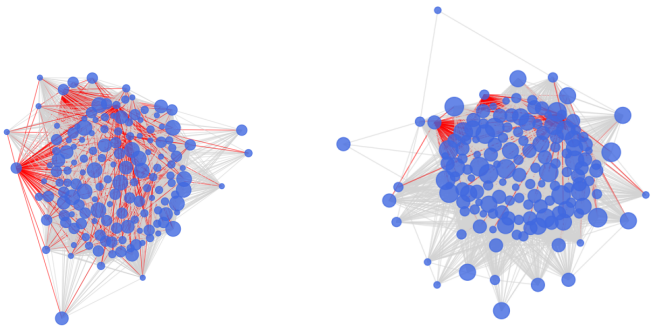


Figure 1. Overview of the proposed framework.

Appendix B. Illustration of the stance-based graph representation



(a) (b)

Figure 2. Examples of constructed graphs for: (a) fake and (b) real news. Grey and red edges correspond to agreements and disagreements, respectively. The node diameter indicates the credibility score of the sender.

Appendix C. Quantitative results using the PHEME dataset

Our experiments follow the established PHEME protocol used in previous studies, where evaluation is performed across different events to avoid topic leakage and to assess cross-event generalization. Table 4 presents comparative results between our method and recent graph-based approaches for rumor detection. We observe that our method consistently outperforms these baselines, primarily due to the incorporation of stance-based representations. This additional analysis positions our approach within the broader rumor detection literature and demonstrates its applicability beyond the original dataset.

Table 4. Comparative results on the PHEME dataset.

Work	**Signal**	**Acc**	**F1**
Dougrez et al. [23]	Web evidence +knowledge graphs	52.30%	48.90%
Wu et al. [12]	Text + propagation tree + temporal windows	75.84%	71.98%
Farinneya et al. [24]	Tweet text + active Learning	-	78.50%
Ours	Stance-aware graph modeling	82.02%	80.64%