Fuse-T Gated Residual Late Fusion of Text Semantics and Thread Topology for Unseen-Event Rumour Classification in Conversational Reply Graphs

Aleksandar Stanković
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
stankovic.sv25.2022@uns.ac.rs
ORCID: 0009-0003-5238-7251

***Abstract*—Rumours on social media spread rapidly during breaking events, while early evidence is often sparse, noisy, and highly event-specific. Text-only classifiers can overfit to keywords and writing style tied to particular events, while structure-only propagation models struggle when discussion trees are small. This paper presents *Fuse-T*, a gated residual late-fusion architecture for binary rumour classification on conversational reply graphs. Fuse-T uses a pre-trained language model as a stable semantic backbone and injects graph propagation cues as a controlled additive residual. A learned element-wise gate modulates the injected topology signal and is initialized near zero to avoid covariate shift on the text classifier. We evaluate Fuse-T on leave-one-event-out (LOEO) generalization across seven events from the PHEME collection. Across events, Fuse-T improves average Macro-F1 from 62.13% (text-only RoBERTa) and 62.42% (text-attributed GNN baseline) to 65.68%. In an early-detection setting using only the first 10 minutes of replies, Fuse-T retains robust performance (about 63% Macro-F1), while structure-only models degrade substantially.**

***Keywords*—rumour detection, early classification, multi-modal fusion, graph neural networks, GraphSAGE, RoBERTa, PHEME**

I. Introduction

Social-media platforms are a primary channel for real-time updates during crises and breaking news, but they also accelerate the spread of unverified claims before reliable corrections are available [1], [2]. This motivates automated *early* rumour classification systems that operate under sparse evidence and distribution shift.

A common operational view treats each instance as a source post and its evolving conversational reply thread, where reactions and conversational context can provide indirect evidence about veracity [3]-[5]. We focus on *unseen-event* generalization using leave-one-event-out (LOEO) evaluation on event-partitioned conversations from PHEME [6]. This setting is practically relevant because deployed models must handle emerging topics, entities, and communities that differ from training events.

Prior work has relied mainly on two signal families: (i) textual content and (ii) propagation structure. Content-based credibility estimation has a long history [7], and modern transformers provide strong semantic representations [8], [9]. However, cross-event transfer remains difficult when models latch onto event-specific lexical correlates (names, locations, hashtags), and social-media language can further deviate from generic pre-training [10]. In parallel, propagation-aware models exploit reply-tree dynamics [11]-[15], but purely structural evidence is often weak in the earliest minutes of a thread [16], [17].

A natural response is to fuse text and topology. Yet many approaches perform *early fusion* by attributing high-dimensional text embeddings to nodes and training a GNN end-to-end, which can be brittle under event shift and may destabilize a strong pre-trained text space. Multimodal learning analyses emphasize that fusion design affects robustness and training stability [18]. Motivated by residual learning and gated fusion [19], [20], we propose *Fuse-T*, a gated residual *late-fusion* architecture that keeps a text encoder as the semantic backbone and injects topology as a controlled additive residual.

Contributions. We make three practical contributions:

- A gate-controlled residual late-fusion module that starts from a stable text-only decision boundary and learns when to inject topology.
- A lightweight, reproducible propagation feature set for conversational reply graphs compatible with inductive GraphSAGE encoders [21].
- A LOEO evaluation on PHEME [6] with per-event Macro-F1 reporting, plus an early-detection protocol that truncates threads by observation time.

II. Related Work

Rumour detection spans dataset curation and shared tasks, early classification under limited evidence, propagation-structure modeling, and text-structure fusion.

A. Datasets, shared tasks, and evaluation

RumourEval (SemEval) operationalized rumour veracity and stance over conversational threads, encouraging models that use conversational context and updates [3], [4]. PHEME provides event-partitioned breaking-news conversations that support studying generalization under event shift [6]. Surveys summarize the rumour lifecycle and the role of conversational and temporal evidence in social media [5].

B. Early detection and temporal dynamics

Early-stage prediction is challenging because both textual cues and propagation structure are sparse and noisy shortly after the source post. Prior work leverages enquiry patterns and early user reactions [16] and shows that performance varies substantially across observation windows, motivating explicit early-time evaluation [17].

C. Propagation structure models

Propagation signals have long been used for credibility estimation [7]. Deep models exploit reply-tree structure via recurrent encoders [11], tree kernels over propagation patterns [12], recursive/tree-structured networks [13], and attention-based tree architectures [14]. Graph neural networks generalize these ideas using message passing on diffusion graphs, including bi-directional formulations [15], but structure-only models can underperform early when reply graphs are small [17].

D. Text encoders and fusion

Transformers are strong text backbones [8], [9], and tweet-adapted pre-training can further improve social-media semantics [10]. Many text-structure systems perform early fusion by attaching node text embeddings and training a GNN end-to-end; while effective in-distribution, this can be brittle under event shift and may amplify noise from sparse early structure. Fuse-T instead follows a conservative late-fusion strategy: topology is projected into the text latent space and injected as a gated residual, aligning with broader multimodal fusion findings [18] and gated/residual design patterns [19], [20].

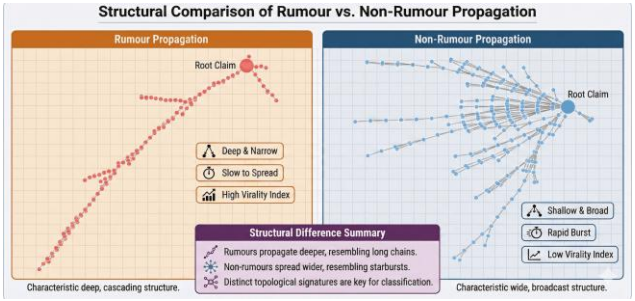

Fig. 2. Illustrative propagation patterns. Rumour threads may form deeper cascades, while non-rumours often resemble shallower broadcast structures. These are tendencies rather than universal rules; Fuse-T uses gating to exploit such cues when they are present

III. Dataset And Task

A. Conversational Reply Graphs

Each discussion thread is represented as a directed reply graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes are posts and edges follow parent→child reply relations. We focus on binary classification: *rumour vs. non-rumour*. Threads are drawn from the PHEME collection of breaking-news conversations. We use a processed leave-one-event-out (LOEO) split over seven events: charliehebdo, ferguson, germanwings, gurlitt, ottawashooting, putinmissing, sydneysiege.

B. Text Construction and Normalization

For text-only and fused models, we encode the root post and optionally concatenate the earliest k replies ordered by length (e.g., 256 subword tokens) and apply standard RoBERTa tokenization. To reduce user- and event-specific leakage, we lower-case and strip URLs and user mentions during preprocessing.

C. Unseen-Event Evaluation (LOEO)

We evaluate generalization with leave-one-event-out (LOEO) splitting: for each fold, all threads from one event are held out for testing while training and validation use the remaining events. This protocol measures robustness to event shift, a central requirement for real-world deployment.

D. Early Detection Setting

To study early classification, we optionally truncate each reply graph to only include posts within the first T minutes from the root (plus the root). If too few nodes remain, we fall back to the earliest K replies. Figure 1 illustrates the gap between a fused model and a structure-only baseline under increasing observation time.

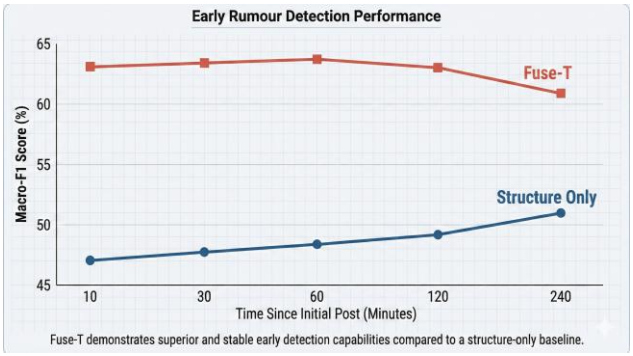

Fig. 1. Early rumour detection (Macro-F1) as a function of time since the initial post. Fuse-T stays stable while a structure-only baseline improves slowly and remains lower.

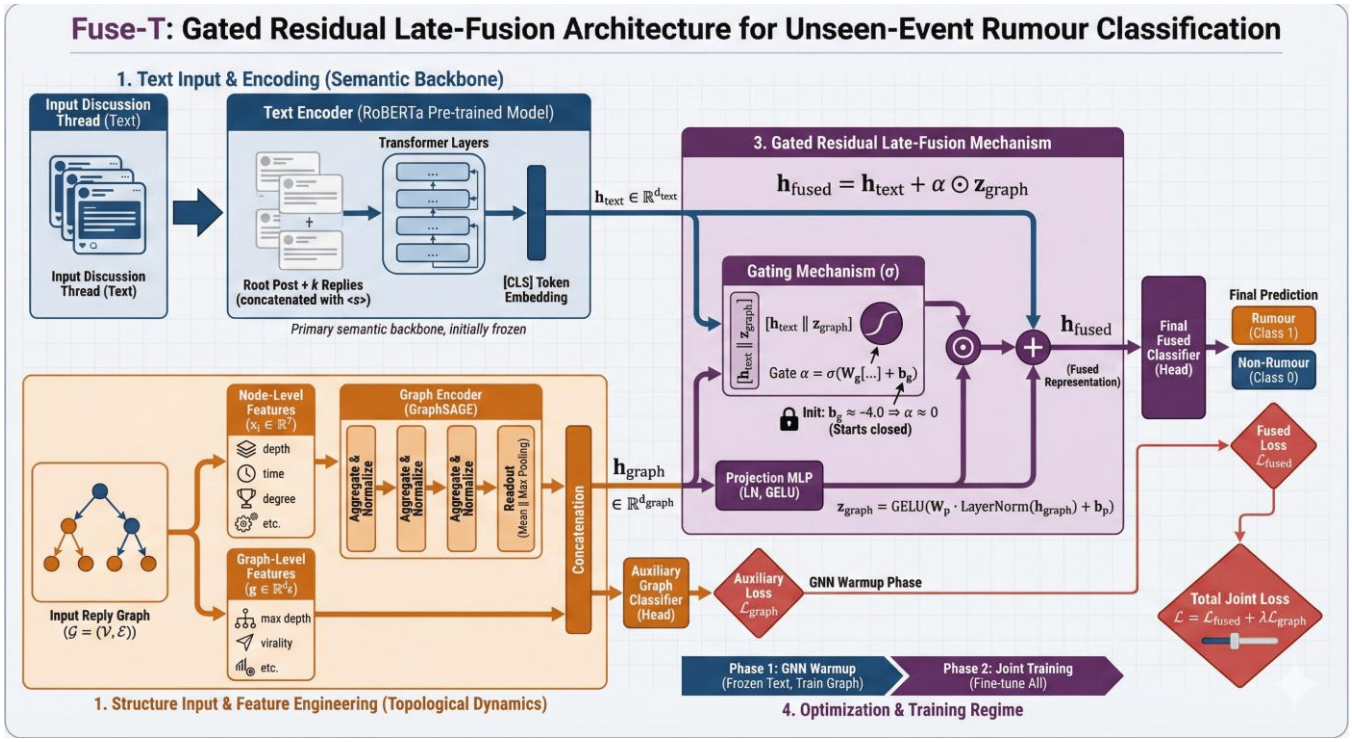

Fig. 3. Overview of Fuse-T. Text and topology are encoded independently (RoBERTa and GraphSAGE). The topology embedding is projected into the text space and injected as a gated residual, initialized near zero to avoid destabilizing the text classifier.

IV. METHOD

Figure 3 summarizes the full Fuse-T pipeline. Fuse-T treats text semantics as the primary backbone and uses topology as an auxiliary, gate-controlled residual.

A. Text Encoder

Given thread text x (root post, optionally concatenated with the earliest k replies), a RoBERTa encoder produces a [CLS] representation $h_{text} \in \mathbb{R}^d$. A linear head maps h_{text} to class logits.

B. Graph Encoder and Propagation Features

We compute propagation-only structural features per node: (i) depth from root, (ii) log time since root, (iii) in-degree and out-degree, (iv) log subtree size, (v) root and leaf indicators, and (vi) simple local branching statistics. A GraphSAGE encoder aggregates these features over the reply graph and produces a graph embedding $h_{graph} \in \mathbb{R}^g$. using a readout (mean, mean+max, or root embedding).

To capture thread-level dynamics, we also use lightweight graph-level features (g_feat), such as node count, maximum depth, average branching factor, degree dispersion (e.g., a Gini proxy), time span, and a coarse temporal growth histogram. These features can help when node features are sparse early on.

C. Gated Residual Late Fusion

Fuse-T aligns modalities by projecting the graph embedding into the text latent space and injecting it as a residual:

$$h_{fused} = h_{text} + \alpha \odot z_{graph}$$

where $z_{graph} = \phi(h_{graph}) \in \mathbb{R}^d$ is a projection MLP and $\alpha \in (0,1)^d$ is an element-wise gate:

$$\alpha = \sigma(W[h_{text}\,;\, z_{graph}] + b)$$

We initialize the gate bias b to a negative value so that $\alpha \approx 0$ initially. Intuitively, Fuse-T starts as a text model and only gradually opens the topology path if it improves validation loss. For fused inference, logits are computed from h_{fused} using the same text classifier head.

D. Training Schedule and Objective

We use a short warmup for the GNN, then train the fusion module and optionally unfreeze the text encoder. Let $\mathcal{L}_{fuse}$ be cross-entropy over fused logits and $\mathcal{L}_{graph}$ a graph-only auxiliary loss. We optimize:

$$\mathcal{L} = \mathcal{L}_{fuse} + \lambda \mathcal{L}_{graph}$$

where λ is a small auxiliary weight. This auxiliary objective stabilizes the GNN encoder so that the fusion module receives a meaningful topology signal.

V. BASELINES

We compare Fuse-T to three baselines.

- **Text (RoBERTa):** thread-level classification from root text (optionally with k replies).
- **GNN (GraphSAGE):** propagation-only GraphSAGE over structural features plus graph-level g_feat.
- **TAG-GNN:** text-attributed GNN, concatenating per-node RoBERTa [CLS] embeddings with structural features (early fusion).

VI. EXPERIMENTAL SETUP

Implementation uses PyTorch, HuggingFace Transformers, and PyTorch Geometric. We report Macro-F1 (primary) and accuracy on each LOEO test fold. Unless otherwise stated, we use RoBERTa-base and GraphSAGE with 2 to 3 layers, hidden dimension 128 to 256, and dropout 0.2 to 0.3. Optimization uses AdamW with separate learning rates for text, graph, and fusion modules. Class imbalance is handled with train-set class weights.

To make hyperparameter choices explicit, Table I summarizes the configuration ranges used in our runs.

TABLE I. HYPERPARAMETER RANGES USED ACROSS LOEO FOLDS.

Component	Setting (range)
Text encoder	RoBERTa-base; max length 128-256
Graph encoder	GraphSAGE layers 2-3; hidden 128-256
Readout	mean / mean+max / root
Dropout	0.2-0.3
Optimizer	AdamW; weight decay 0.01
LR (text)	1e-5-3e-5 (optional unfreeze)
LR (graph)	5e-4-2e-3
LR (fusion)	2e-4-1e-3
Gate bias init	$b \in [-6, -2]$ (starts closed)
Aux Weight	$\lambda \in [0.1, 0.5]$

VII. RESULTS

A. Unseen-Event Performance

Table II reports Macro-F1 (%) per held-out event. On average, Fuse-T achieves 65.68% Macro-F1, outperforming the text-only baseline (62.13%) and TAG-GNN (62.42%). Notably, Fuse-T improves substantially on the challenging gurlitt event, where pure text and TAG-GNN degrade.

TABLE II. MACRO-F1 (%) ACROSS HELD-OUT PHEME EVENTS (LOEO).

Event	Text	GNN	TAG-GNN	Fuse-T
charliehebdo	**78.24**	56.54	76.95	75.91
ferguson	52.86	50.80	50.97	**54.93**
germanwings	71.25	47.57	**71.83**	71.24
gurlitt	37.27	48.30	35.51	**49.37**
ottawashooting	71.08	56.72	71.28	**73.68**
putinmissing	53.68	49.89	58.46	**60.99**
sydneysiege	70.50	57.57	71.94	**73.64**
Avg.	62.13	52.48	62.42	**65.68**

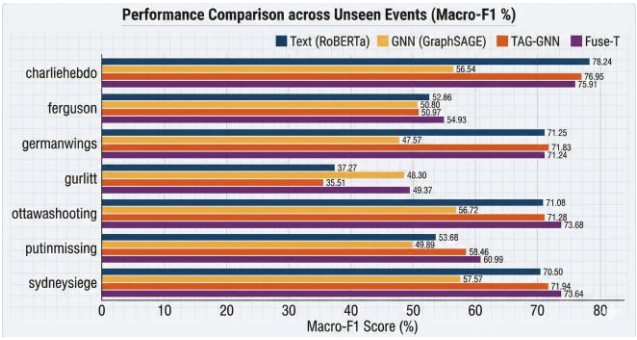

Fig. 4. Macro-F1 (%) per unseen event (LOEO). Fuse-T is competitive or best on most events, with large gains on the most difficult fold (gurlitt).

B. Fusion and Gating Ablations

Table III isolates the effect of the fusion mechanism. The gated residual injection with a closed-at-initialization bias (bias=-4) performs best on average, improving Macro-F1 over (i) removing the gate (always injecting topology), (ii) initializing the gate unbiased (bias=0), and (iii) a concatenation fusion baseline. This supports the design goal: starting from a stable text decision boundary and only injecting topology when beneficial yields more robust unseen-event generalization.

TABLE III. FUSION/GATING ABLATIONS ON LOEO FOLDS (MEAN ± STD ACROSS HELD-OUT EVENTS). VALUES ARE IN PERCENTAGE POINTS.

Ablation	Macro-F1 (↑)	Acc. (↑)
Residual + Gate (bias=-4)	**66.2 ± 12.9**	**70.7 ± 9.4**
Residual (no gate)	63.6 ± 9.7	64.8 ± 7.3
Residual + Gate (bias=0)	65.3 ± 10.9	66.2 ± 9.0
Concat fusion	61.6 ± 12.6	63.8 ± 8.4

C. Where Does Fusion Help?

To make gains easier to interpret, Table IV reports event-wise improvements of Fuse-T over (i) text-only and (ii) TAG-GNN. While Fuse-T slightly underperforms text on charliehebdo (a text-friendly event), it provides meaningful positive gains on most other events, with the largest improvements on gurlitt and putinmissing.

TABLE IV. EVENT-WISE MACRO-F1 GAINS OF FUSE-T (%) OVER TEXT-ONLY AND TAG-GNN.

Event	Δ vs. Text	Δ vs. TAG-GNN
charliehebdo	-2.33	-1.04
ferguson	2.07	3.96
germanwings	-0.01	-0.59
gurlitt	**12.10**	**13.86**
ottawashooting	2.60	2.40
putinmissing	7.31	2.53
sydneysiege	3.14	1.70
Avg.	3.55	3.26

D. Early Detection

With early truncation to the first 10 minutes of a thread, propagation-only models are sensitive to sparse trees. Fuse-T remains comparatively robust (about 63% Macro-F1 in the early setting) by relying on the stable text backbone and injecting topology only when the learned gate deems it beneficial.

VIII. DISCUSSION

Fuse-T improves average Macro-F1 under LOEO evaluation compared to both a text-only transformer and a text-attributed GNN baseline, with the largest gains concentrated on the most challenging held-out events. This pattern matches the intended role of topology as *auxiliary evidence*: when lexical cues transfer well, a strong text backbone can already perform competitively, while structural signals become most useful when cross-event text generalization degrades [5], [15].

Ablations support the design choice of *gated residual injection*. Removing the gate or starting from an unbiased gate increases variance and can reduce mean performance, consistent with the risk that unconstrained fusion amplifies spurious correlations or introduces optimization instability. From a multimodal-learning perspective, Fuse-T behaves as a conservative late-fusion method that begins near a strong unimodal solution and learns when the auxiliary path is predictive [18]. The residual form aligns with additive refinement in residual learning [19], while the gate provides explicit control over cross-modal influence [20].

Early-detection results highlight a further benefit: when reply graphs are small, structure-only models have limited evidence and typically improve slowly as the observation window grows [17]. Fuse-T remains more stable because it can rely on semantics early and only inject topology when the learned gate indicates that structural cues are informative, in line with findings that early cues are often weak and noisy [16].

Finally, several extensions follow naturally from established benchmarks and practice. Incorporating explicit stance modeling could strengthen evidence aggregation in conversational settings [3], [4], [22]. Domain-adaptive pre-training for social-media text may further reduce event-specific artifacts and improve cross-event transfer [10]. More expressive temporal encodings could also improve reliability when early propagation structure is ambiguous [17].

IX. CONCLUSION AND FUTURE WORK

We presented Fuse-T, a gated residual late-fusion model for unseen-event rumour classification on conversational reply graphs. Across seven LOEO PHEME events, Fuse-T improves average Macro-F1 to 65.68% and remains robust in an early-detection setting. Future work includes richer temporal encodings, uncertainty-aware gating, and evaluation on additional datasets and platforms.

ACKNOWLEDGMENT

We gratefully acknowledge computational support from Xinming Wang at the Institute of Automation, Chinese Academy of Sciences (CASIA).

REFERENCES

[1] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga, "Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval), 2017, pp. 69–76.

[4] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, and K. Bontcheva, "Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours," in Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval), 2019, pp. 845–854.

[5] A. Zubiaga, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," ACM Computing Surveys, vol. 51, no. 2, pp. 32:1–32:36, 2018.

[6] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," PLOS ONE, vol. 11, no. 3, p. e0150989, 2016.

[7] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in Proceedings of the 20th International Conference on World Wide Web (WWW), 2011, pp. 675–684.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[10] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," in Proceedings of EMNLP (Systems Demonstrations), 2020, pp. 9–14.

[11] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 3818–3824.

[12] J. Ma, W. Gao, and K.-F. Wong, “Detect rumors in microblog posts using propagation structure via kernel learning,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, 2017, pp. 708–717.

[13] J. Ma and W. Gao, “Rumor detection on Twitter with tree-structured recursive neural networks,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, 2018, pp. 1980–1989.

[14] J. Ma and W. Gao, “Debunking rumors on Twitter with tree transformer,” in Proceedings of the 28th International Conference on Computational Linguistics (COLING), 2020, pp. 5455–5466.

[15] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, “Rumor detection on social media with bi-directional graph convolutional networks,” in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 34, no. 01, 2020, pp. 549–556.

[16] Z. Zhao, P. Resnick, and Q. Mei, “Enquiring minds: Early detection of rumors in social media from enquiry posts,” in Proceedings of the 24th International Conference on World Wide Web (WWW), 2015, pp. 1395–1405.

[17] S. Kwon, M. Cha, and K. Jung, “Rumor detection over varying time windows,” PLOS ONE, vol. 12, no. 1, p. e0168344, 2017.

[18] T. Baltrus˘aitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[20] J. Arevalo, T. Solorio, M. Montes-y Go´mez, and F. A. Gonza´lez, “Gated multimodal units for information fusion,” arXiv preprint arXiv:1702.01992, 2017.

[21] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 1024–1034.

[22] E. Kochkina, M. Liakata, and I. Augenstein, “Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm,” in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval), 2017, pp. 475–480.