On Fake News Detection with LLM Enhanced Semantics Mining

Anonymous ACL submission

Abstract

Large language models (LLMs) have emerged as valuable tools for enhancing textual features in various text-related tasks. In this paper, we assess the effectiveness of news embeddings from ChatGPT for detecting fake news and showcase that despite their initial performance slightly surpassing the pre-trained BERT model, they still lag behind the state-of-the-arts. This shortfall is attributed to the reliance on tokenized training text, which misses the complex narratives and subtleties that are crucial for identifying fake news. To capture these nuances, we probe the high-level semantic relations among the news pieces, real entities, and topics, which are modeled as a heterogeneous graph with nodes denoting different items and the relations are represented as edges. We then propose a Generalized PageRank model and a consistent learning criteria for mining the local and global semantics centered on each news piece through the adaptive propagation of features across the graph. Our model shows new state-of-the-art performance on five benchmark datasets and the effectiveness of the key ingredients is supported by extensive analysis. Our code is available at https://github.com/LEG4FD/LEG4FD.

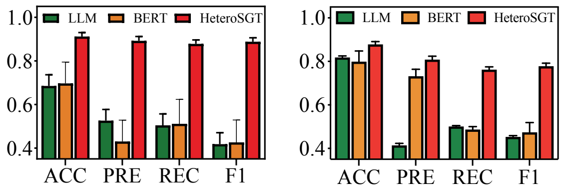
1 Introduction

The ubiquity of fake news on social media poses a significant threat to public discourse and societal well-being (Prieur et al., 2023; Chen et al., 2023). As to alleviate the far-reaching consequences, many fake detection methods probe the information dissemination process or social structure (Mehta et al., 2022; Hu et al., 2021; Su et al., 2023) to detect fake news. Unfortunately, despite the impressive detection performance, their applicability is substantially constrained when the social context is unavailable or incomplete due to the evolving nature of social networks and data privacy concerns (Zhou and Zafarani, 2020; Zhang and Ghorbani, 2020). Facing limited access to social context, other text-mining methods (Yang et al., 2016; Zhang et al., 2024) investigate the intricacies of news content to uncover hierarchical textual semantics (e.g., sentence and document level semantics) and formulate fake news detection as a classification problem, using only textual content from the social media.

Following the latter approach, in which the news embeddings are critical in providing a discriminatory description of authentic and fake news, we are propelled to enhance them with Large Language Models (LLMs), which have been renowned for their remarkable capabilities in language understanding, and context modeling (Thota et al., 2018; Zhao et al., 2023; Li et al., 2023). A fundamental question that guides our research in this under-explored realm is, *"Are the LLM output news embeddings effective for fake news detection?"*

To this, we conducted a preliminary study by comparing the detection performance of an MLP classifier trained using news features extracted from LLM[1], BERT (Devlin et al., 2018) and HeteroSGT (Zhang et al., 2024), respectively. From the results depicted in Fig. 1, we found that the LLM extracted features slightly outperform those from BERT, but significantly behind HeteroSGT. On the one hand, such an undermined performance of LLM and BERT highlights that the embedding-based enhancement (Li et al., 2023), which generates initial embeddings following $\boldsymbol{x} = f_{\text{LM}}(t)$, is insufficient to encapsulate the nuanced semantics for effective fake news detection. Here, for brevity, we use t to denote the textural content of a particular news piece. On the other hand, since HeteroSGT also employs Transformer (Vaswani et al., 2017) as the backbone but investigates the high-level semantics among news, entities, and topics for fake news detection, it outperforms LLM and

[1] https://platform.openai.com/docs/api-reference/embeddings



(a) ReCOVery (b) MC Fake

Figure 1: A comparison between fake news detection performance on two datasets w.r.t. **acc**uracy, **pre**cision, **rec**all and **F1** scores.

Method	Source of Features			Semantics		Unlabeled Data
	Social Context	News Text	Other Sources	Local	Global	
HAN	✗	✓	✗	✓	✗	✗
TextGCN	✗	✓	✗	✓	✗	✗
DualEmo	Comments	✓	✗	✓	✗	✗
UsDeFake	Propagation Network	✓	✗	✓	✗	✗
HGNNR	✗	✓	Knowledge Graph	✓	✗	✗
HeteroSGT	✗	✓	✗	✓	✓	✗
LEG4FD(Ours)	✗	✓	✗	✓	✓	**CR**

Table 1: Overview of fake news detection methods. Comparisons are made regarding the sources of features for fake news detection, the semantics each method explores, and how they enforce the learning on unlabeled data.

BERT, which only consider the lexical semantics between tokens. As a result, we raise two further sub-problems to the incorporation of LLM for better detection performance:

- *How can we apply LLM to high-level news semantics exploration?* Though LLMs are powerful in language analysis, the keys for high-level semantics exploration are to extract *distinct entities with real meaning* and *the narratives*.
- *How can we achieve fine-gained news representations using LLM-derived semantic information?* Aggregating semantic information of individual news pieces (Thota et al., 2018; Zhang et al., 2024) focuses solely on local semantics and overlooks the valuable global semantics across news. It is crucial to incorporate the intricate details of individual articles and the broader contextual insights from all news pieces.

To address ***sub-problem 1***, in addition to prompting LLM for entity extraction, we first propose a refined topic model that summarizes news topics through LLM-generated embeddings and then construct a heterogeneous graph to model the relations between news pieces, entities, and topics. Consequently, the complex news narratives can be described by the news embeddings and the edges between other nodes. For ***sub-problem 2***, we propose to apply short- and long-scale feature propagation centered on news nodes to generate fine-gained news representations that capture both the local and global semantics. Empowered by the two scales of feature propagation, we further introduce a consistency learning criteria to involve unlabeled news for training. Our major contributions are: **1)** We evaluate different news feature enhancement strategies leveraging LLMs, uncovering two fundamental problems that should be addressed to incorporate LLMs for advancing the detection of fake news; **2)** We introduce an LLM-enhanced topic model and devise potent prompts for querying LLMs. Our method, LEG4FD, models the intricate semantics among news pieces, entities, and topics within a heterogeneous graph, which facilitates the exploration of both local semantics surrounding individual news and global semantics spanning across the dataset; **3)** Our proposed feature propagation model not only captures the local and global news semantics on label news, but also allows a flexible consistency regularization on unlabeled data for refining the news representation; and **4)** Through extensive experiments on five real-world datasets, our method shows new state-of-the-art performance. The effectiveness of our design choices is validated through further case studies.

2 Related Work

2.1 Fake News Detection

Current investigations into fake news detection can be categorized into *content-based* and *graph-based* methodologies, in terms of their focus on specific aspects of news articles for feature mining. Specifically, the content-based methods concentrate on analyzing the textual content of news articles, extracting linguistic, syntactic, stylistic, and other textual features to differentiate between genuine and fake news. For example, Horne and Adali (2017) and Kaliyar et al. (2021) analyzed the language styles to distinguish between fake and real news while Yang et al. (2016) introduced a dual-attention model to explore hierarchical news semantics. Other works also explored the incorporation of supplementary textual information, such as comments (Shu et al., 2019; Rao et al., 2021), and emotion signals (Zhang et al., 2021), to further improve detection capabilities. These content-based methods strive to explore diverse textual features associated with each single article to identify their authenticity. However, the detection performance is compromised when fake news is specially fabricated to mimic the words and language styles of genuine news, which inherently necessitates the need to explore higher-level semantics, such as the relations among news, real entities, and topics that are explored in this paper.

Moving beyond the content-based methods, graph-based methods explicitly model and learn potential structures, such as word-word relations (Yao et al., 2019; Linmei et al., 2019), news dissemination graphs (Ma et al., 2018; Bian et al., 2020), and social structure (Su et al., 2023; Dou et al., 2021). Concrete examples under this category include: Yao et al. (2019) which first constructed a weighted graph using the words within the news content and then applied the graph convolutional network (GCN) for classifying fake news; Linmei et al. (2019) that built a similar graph but employed a heterogeneous graph attention network for classification (Linmei et al., 2019); and Bian et al. (2020) which employed recurrent neural networks and bidirectional GCN to capture the new features from their propagation process. There are other works that model the relations between news and users (Su et al., 2023; Dou et al., 2021), or even news and external knowledge sources (Hu et al., 2021; Dun et al., 2021; Xu et al., 2022; Xie et al., 2023; Wang et al., 2018) to complement fake news detection. Despite their progress, the reliance on supplementary sources poses a notable challenge in their applicability, and even when this auxiliary information is available, the associated computational costs remain an additional hurdle. For clarity, we summarize our work and the existing methods in Table 1.

2.2 Language Models for Feature Mining

Large Language Models (LLMs) such as GPT (Radford et al., 2018), and Pre-trained language models like BERT (Devlin et al., 2018) have emerged as powerful tools for feature mining due to their remarkable adaptability in language understanding, sentiment analysis, machine translation, and text classification (Min et al., 2023; Liu et al., 2023; Wu and Ong, 2021). The utilization of LMs for feature mining aims to enrich the embeddings of input texts. And the most straightforward application is to feed the output features for training machine learning models that are tailored to specific tasks, such as time series analysis and graph learning (Jin et al., 2023; Li et al., 2023).

To get more specific information and further enrich the textual representations, more advanced methods prompt LLMs to generate supplementary content, such as explanations, related knowledge, and background information (Min et al., 2023). This additional content is then combined with the original texts for downstream modeling. For example, He et al. (2023) took a pre-trained language model to encode both text data and LLM-generated explanations as initial node embeddings for better text-attributed graph representation learning. Li et al. (2022) explored the potential of the explanations generated by LLMs to improve the reasoning capability of relatively small language models. In summary, LMs showcase their potential for advancing various natural language processing-related tasks, and this paper targets utilizing LLMs for news semantics modeling by mitigating the two prior recognized sub-problems.

3 Methodology

3.1 Preliminaries

DEFINITION 1. **Heterogeneous Graph.** A heterogeneous graph $\mathcal{HG} = \{\mathbb{V}, \mathbb{L}, \mathbb{X}\}$ models the intricate relations (in $\mathbb{L}$), among diverse types of instances in $\mathbb{V}$. For fake news detection, our node set $\mathbb{V} = \{n_i\}_{i=0}^{|\mathbb{N}|} \cup \{e_i\}_{i=0}^{|\mathbb{E}|} \cup \{t_i\}_{i=0}^{|\mathbb{T}|}$ comprises three distinct types of nodes: *news nodes* ($\mathbb{N}$), *entity nodes* ($\mathbb{E}$) and *topic nodes* ($\mathbb{T}$). Each link/edge in $\mathbb{L}$ denotes the explicit relation between two nodes. $\mathbb{X} = \{\mathbf{X}^n, \mathbf{X}^e, \mathbf{X}^t\}$ encompasses the feature vectors for all nodes, in which $\mathbf{X}^n \in \mathbb{R}^{|\mathbb{N}|\times d}$ is the news node feature matrix, $\mathbf{X}^e \in \mathbb{R}^{|\mathbb{E}|\times d}$ for entities and $\mathbf{X}^t \in \mathbb{R}^{|\mathbb{T}|\times d}$ for topics.

DEFINITION 2. **Fake News Detection.** In this paper, we define fake news detection as to learn a model $\mathcal{M}(\cdot)$ using the text of both labeled news $(\mathbb{N}^L, \mathbb{Y}^L)$ and unlabeled news $\mathbb{N}^U$, to infer the labels of the unlabeled news, $\hat{\mathbb{Y}}^U$. For a particular news n_i, its label $y_i \in \mathbb{Y}^L \cup \mathbb{Y}^U$ is 1 if the news is fake, and 0 if it is authentic.

3.2 LLM-Enhanced Semantics Modeling

News articles naturally encompass various *entities* with real meaning, such as people, locations and organizations, and usually focus on specific *topics*. These named entities and topics comprise rich high-level semantic information and narratives about news articles, which are crucial for identifying the nuance of fake news. Driven by our preliminary study results, as depicted in Fig. 1, we further investigate LLMs, particularly ChatGPT, to address our devised ***sub-problem 1*** as follows.

Entity Extraction. For news entity extraction, we prompt the LLM following Table 2 for identi-

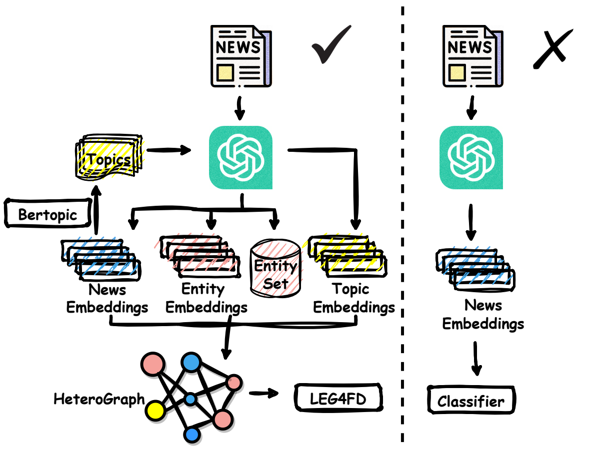


Figure 2: Heterogeneous graph construction.

fying specific entities in all news pieces including persons, dates, locations, organization, and miscellaneous entities[2].

News and Entity Embedding. We obtain the news embeddings and entity embeddings by directly querying the API provided by OpenAI[3] to encode the corresponding content. The resulting news embeddings are processed as $\mathbf{X}^n$, and the entity embeddings are stored in $\mathbf{X}^e$.

Topic Modeling. Motivated by Zhang et al. (2024) in exploring news on the same topic, we acknowledge the value of topics not only for summarizing the news focus and linking different news, but also for exploring the relation between a target news and entities in another news, as supported by the empirical results in Sec. 4.3. For involving the topic information for fake news detection, we adopt Bertopic (Grootendorst, 2022) to derive the topics involved in all news, which typically outputs the topic words and the corresponding weights for each topic. We then feed the topic words into the API call to extract their embeddings from LLM and formulated the embedding of each topic as the weighted sum of topic words within it following:

$$\boldsymbol{x}_i^t = \sum_{j \in \mathcal{B}(t_i)} w_{j,t} \boldsymbol{h}_j; \quad \boldsymbol{x}_i^t \in \mathbf{X}^t, \qquad (1)$$

where $\mathcal{B}(t_i)$ is the topic word list output by Bertopic, $w_{j,t}$ is the corresponding weight of word j to topic t_i, and $\boldsymbol{h}_j$ is the topic word embedding from LLM.

[2]Notably, we only input the widely-used and public available datasets for querying the LLM in case of any privacy and ethical concerns.

[3]https://api.openai.com

PROMPT:

 # Task

Extract the following entities from the given news article:
1. PERSON: *Person Definition.* 2. DATE: *DATE Definition.*
3. LOC: *LOC Definition.* 4. *ORG: ORG Definition.*
5. *MISC: MISC Definition.*
Return the results in a dictionary with corresponding keys.

 # Examples

Example 1: "The iPhone, created by Apple Inc., was released on June 29, 2007."
Output1: "PERSON": ["None"], "DATE": ["June 29, 2007"], "LOC": ["None"], "ORG": ["Apple Inc."], "MISC": ["iPhone"]
Examples 2: ...
Output2: ...

 # Input News Article

Given news article: < *The SpaceX CEO, Elon Musk, announces ambitious plans to build a self-sustaining underwater city on Mars by Dec 2030 ...* >

CHATGPT:

"PERSON": ["Elon Musk", ...], "DATE": ["Dec 2030", ...], "LOC": ["Mars", ...], "ORG": ["SpaceX", ...], "MISC": ["CEO", ...]

Table 2: Prompt for entity extraction.

For replication purposes, we detail the practical settings in entity extraction, embedding, and topic modeling in Sec 4, accompanied by the in-depth analysis of their empirical impact.

Heterogeneous Graph Construction. Given the news pieces, entities, topics, and their corresponding embeddings, we then follow Definition 1 and construct a heterogeneous graph $\mathcal{HG}$, in which we consider two types of explicit relations: <news, 'contains', entity> and <news, 'focuses on', topic>.

In summary, we construct a heterogeneous graph, $\mathcal{HG}$, to capture: 1) *high-level relationships* among news items, entities, and topics, represented as edges; and 2) *sentence/document-level narratives* encapsulated within the embeddings of news items, entities, and topics, denoted by $\mathbb{X}$. This approach addresses our recognized ***sub-problem 1*** and facilitates a thorough examination of local semantics around each news item, exemplified by the 1-hop or 2-hop subgraphs centered on news nodes in $\mathcal{HG}$, as well as global semantics across broader ranges, all empowered by LLM.

3.3 Generalized Feature Propagation

Given $\mathcal{HG}$, we propose to learn fine-grained news representations by encapsulating the valuable information in entities, topics and other similar news that share common topics or entities. It is worth noting that we highlight the significance of exploring these high-level semantics not only because of the preliminary results reported in Fig. 1, but also regarding the consensus that fake news carries false knowledge about real entities on a particular topic (Zhou and Zafarani, 2020). Therefore, we take news, entities, and topics into account so as to distinguish the nuances of fake news.

We propose to use Generalized PageRank (GPR) for propagating the features of entities, topics, and other news pieces to the target, by simply learning a weighing scalar for each propagation step. To be specific, we first apply a two-layers MLP, $f_{\boldsymbol{\theta}}(\cdot)$, and project the news, entities and topics' features into the same space following $\mathbf{H} = f_{\theta}(\mathbf{X})$, and $\mathbf{X} = [\mathbf{X}^{n\top}, \mathbf{X}^{e\top}, \mathbf{X}^{t\top}]^\top$ is the vertical stack of the three feature matrices. As to facilitate feature propagation, we then unify the index of all three types of nodes based on their index in $\mathbf{X}$ and transform the heterogeneous graph structure into a homogeneous adjacency matrix, $\mathbf{A}$, with regard to the edges in $\mathcal{HG}$ and by adding self-loops. A particular element $\mathbf{A}_{[i,j]} = 1$ if there exists an edge between nodes i and j in $\mathcal{HG}$.

With the projected node features $\mathbf{H}$ and adjacency matrix $\mathbf{A}$, we can promptly propagate the features following:

$$\mathbf{H}^s = \mathbf{P}\mathbf{H}^{s-1}, \tag{2}$$

where s denotes the propagation step, $\mathbf{H}^0 = \mathbf{H}$, and $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ is the row normalized adjacency matrix given the diagonal degree matrix $\mathbf{D}$. Then, the target news representations is formulated as the weighted sum of the propagated features in S steps, given by:

$$\mathbf{Z} = \sum_{s=0}^{S} w_s \mathbf{H}^s, \tag{3}$$

where w_s is a learnable weight corresponding to step s and the value can be either positive or negative, indicating how the information at a particular step contributes to the prediction. Thus, the learned news representations comprise the high-level semantics information within S steps, and the probabilities of a news piece being authentic or fake is predicted as $\boldsymbol{p}_i = \mathrm{softmax}(\boldsymbol{z}_i)$,which can be directly applied to enforce the learning of $\boldsymbol{\theta}$ and $\boldsymbol{w}$ using the cross-entropy loss on labeled news. However, this only preserves the semantics within a particular scale S.

3.4 Global and Local Semantics Mining

During feature propagation, a larger step allows the exploration of global semantics across $\mathcal{HG}$ since neighbors across broader ranges are involved, while a smaller step stresses more to the local semantics between the target news piece and its highly related entities, topics and news. Both scales of semantics offer complementary perspectives on the target news and we can firmly apply two divergent scale values s_g and s_l to encode the **g**lobal and **l**ocal semantics into news embeddings, respectively. By setting a small step s_l (e.g., 2) and a larger step s_g (e.g., 20), we can obtain two representations, $\boldsymbol{z}_i^l \in \mathbf{Z}^l$ and $\boldsymbol{z}_i^g \in \mathbf{Z}^g$ for each news pieces following Eq.(3). Indeed, these representations can be viewed as two divergent augmentations of the news pieces from the perspective of data augmentation, and we enforce the cross-entropy loss on both views to train the model on the labeled news, which is to minimize:

$$\mathcal{L}_{sup} = \frac{1}{|\mathbb{N}^L|} \sum_{i \in \mathbb{N}^L} \left[\mathcal{L}_{ce}(\boldsymbol{p}_i^l, y_i) + \lambda_g \mathcal{L}_{ce}(\boldsymbol{p}_i^g, y_i) \right], \tag{4}$$

where $\boldsymbol{p}_i^l$ and $\boldsymbol{p}_i^g$ are the predictions made upon the news embeddings $\boldsymbol{z}_i^l$ and $\boldsymbol{z}_i^g$, respectively. λ_g balances the contributions of the local and global semantics.

3.5 Consistency Regularization on Unlabeled News

Since our learned news representations already comprise the global and local semantics, we further explore regularization signal from unlabeled data to make consistent predictions upon $\mathbf{Z}^l$ and $\mathbf{Z}^g$. Our proposed regularization term comprises two dependent ingredients: 1) prototype estimation; and 2) consistency loss between the predictions. Specifically, the prototype estimation is to align the predictions $\boldsymbol{p}_i^l$ and $\boldsymbol{p}_i^g$ on each node, which follows:

$$\overline{\boldsymbol{p}_i} = (\boldsymbol{p}_i^l + \lambda_g \boldsymbol{p}_i^g)/2. \tag{5}$$

Then, we define the consistency loss on unlabeled news as the overall prediction divergence between the prototype and two views following:

$$\mathcal{L}_{con} = \frac{1}{2|\mathbb{N}^U|} \sum_{i \in \mathbb{N}^U} \left[\mathcal{D}(\overline{\boldsymbol{p}_i} || \boldsymbol{p}_i^l) + \lambda_g \mathcal{D}(\overline{\boldsymbol{p}_i} || \boldsymbol{p}_i^g) \right], \tag{6}$$

where $\mathcal{D}(\cdot)$ measures the KL-divergence.

Notably, our model design features an end-to-end optimization of both the scale weights ($\boldsymbol{w}$) and the MLP parameters ($\boldsymbol{\theta}$). The inclusion of this consistency loss not only regularizes the propagation of more valuable features into new representations - capturing both local and global semantics effectively; but also enhances the detector's generalization capabilities on unlabeled data.

Dataset	# Fake	# Real	# Total	# Entities
MM COVID	1,290	869	2,159	3,353
ReCOVery	578	1,254	1,832	13,703
MC Fake	2,591	12,435	15,026	150,435
LIAR	1,595	1,346	2,941	4,066
PAN2020	238	243	481	9,740

Table 3: Statistics of datasets.

3.6 Training Objective and Fake News Detection

Combing both the supervised loss and consistency loss, we formulated the overall training objective of our method as:

$$\arg\min_{\boldsymbol{w},\boldsymbol{\theta}} \lambda_{ce}\mathcal{L}_{sup} + \lambda_{cr}\mathcal{L}_{con}, \tag{7}$$

where λ_{ce} and λ_{cr} trades off the training signals from the labeled and unlabeled news. After training, we promptly predict the label of each news as $\hat{y}_i = \arg\max(\overline{\boldsymbol{p}_i})$, where i is classified as fake if $\hat{y}_i = 1$, and as authentic otherwise.

4 Experiment

Evaluation Dataset. Our evaluation datasets cover diverse domains, including health-related datasets (MM COVID (Li et al., 2020) and ReCOVery (Zhou et al., 2020)), a political dataset (LIAR (Wang, 2017)), and multi-domain datasets (MC Fake (Min et al., 2022) and PAN2020 (Rangel et al., 2020)). Notably, the MC Fake dataset includes news articles across politics, entertainment, and health, sourced from reputable debunking websites, such as PolitiFact[4] and GossipCop[5]. Statistics of these datasets are provided in Table 3.

Baselines. We compare LEG4FD against seven representative baselines in text classification and fake news detection, including **textCNN** (Kim, 2014), **textGCN** (Yao et al., 2019), **BERT** (Devlin et al., 2018), **SentenceBERT** (Reimers and Gurevych, 2019), and **HAN** (Yang et al., 2016) that work on word tokens from news text for classification; **HGNNR4FD** (Xie et al., 2023) and **HeteroSGT** (Zhang et al., 2024), which model the high-level news semantics as a graph for fake news detection. We exclude other methods that are reliant on propagation information (Wei et al., 2022; Yang et al., 2022), social engagement (Shu et al., 2019; Zhang et al., 2021), and alternative sources of evidence (Xu et al., 2022; Khattar et al., 2019) to ensure a fair comparison. We also ignore the conventional heterogeneous graph neural networks because HeteroSGT has already demonstrated superior performance over them. A summary of the baselines is provided in Appendix A.1.

Experimental Settings. Throughout the experiment, we fix the dimensionalities of the two MLP layers as 64 and 2, respectively, and employ the Adam optimizer with a learning rate of 0.002 and weight decay $5\mathrm{e}{-4}$. To test the generalizability, we perform 10-fold cross-validation (using a split ratio of 80%-10%-10% for training, validation and test) and report the averaged results (in percentage) along with the standard deviations with regard to five mostly-used metrics: Accuracy (Acc), macro-precision (Pre), macro-recall (Rec), macro-F1 (F1), and the AUC-ROC curve. Detailed hyperparameter settings are provided in Appendix A.2.

4.1 Fake New Detection Performance

Overall Performance. The results summarized in Table 4 and Fig. 4 reveal that our method, LEG4FD, significantly surpasses all baseline models w.r.t. the four evaluation metrics. The performance gaps, which is over 5% on MM COVID and 2% on the rest datasets, affirm the effectiveness of our approach in investigating the LLM-enhanced news semantics solely on the textual content of news. Further comparative analysis with the baseline models yields additional insights:

High-level Semantic Exploration is Pivotal. Despite the effectiveness of traditional classifiers like TextCNN, TextGCN, HAN, BERT, and Sentence-BERT in capturing word-level narratives, they struggle with the relationships among news pieces, entities, and topics, limiting their performance. In contrast, our method, along with HeteroSGT and HGNNR4FD, excels by modeling these high-level semantics in a graph, analyzing the relations and features of news, entities, and topics for enhanced results.

Mining the Global and Local Semantics Results in the Better Performance. While HGNNR4FD and HeteroSGT employ heterogeneous graphs to analyze news, entities, and topics, their performance are deteriorated due to the insufficient exploration of global and local semantics. Specifically, HGNNR4FD focuses on semantics at a specific scale, while HeteroSGT suffers from information loss through random walks. Our method surpasses these issues, efficiently mining global and local semantics with lower computational demands, as detailed in Section 4.4.

[4] https://www.politifact.com
[5] https://www.gossipcop.com

Dataset	TextCNN		TextGCN		HAN		BERT		SentenceBert		HGNNR4FD		HeteroSGT		LEG4FD	
	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre
MM COVID	0.564±0.038	0.484±0.173	0.691±0.160	0.716±0.240	0.829±0.009	0.836±0.007	0.744±0.010	0.705±0.010	0.761±0.004	0.786±0.002	0.732±0.017	0.882±0.016	0.924±0.011	0.918±0.012	**0.974±0.010**	**0.975±0.010**
ReCOVery	0.649±0.002	0.449±0.107	0.733±0.004	0.697±0.183	0.694±0.003	0.435±0.201	0.697±0.003	0.430±0.214	0.687±0.006	0.645±0.167	0.783±0.008	0.771±0.006	0.912±0.010	0.892±0.014	**0.938±0.020**	**0.930±0.018**
MC Fake	0.816±0.004	0.530±0.159	0.697±0.142	0.524±0.173	0.834±0.004	0.444±0.103	0.799±0.005	0.732±0.003	0.828±0.002	0.464±0.006	0.818±0.010	0.456±0.010	0.878±0.012	0.808±0.012	**0.894±0.012**	**0.826±0.015**
LIAR	0.556±0.002	0.447±0.185	0.487±0.039	0.493±0.047	0.559±0.003	0.501±0.005	0.522±0.003	0.522±0.002	0.566±0.002	0.565±0.002	0.544±0.013	0.559±0.009	0.582±0.017	0.579±0.016	**0.678±0.021**	**0.765±0.019**
PAN2020	0.503±0.002	0.309±0.119	0.495±0.032	0.392±0.144	0.494±0.005	0.457±0.135	0.519±0.005	0.541±0.005	0.524±0.005	0.508±0.009	0.690±0.014	0.677±0.14	0.720±0.021	0.731±0.021	**0.771±0.017**	**0.798±0.019**
Dataset	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**	**Rec**	**F1**
MM COVID	0.560±0.004	0.492±0.104	0.694±0.181	0.642±0.245	0.834±0.04	0.838±0.009	0.723±0.112	0.711±0.103	0.730±0.006	0.729±0.006	0.648±0.021	0.755±0.021	0.912±0.012	0.916±0.012	**0.973±0.009**	**0.973±0.010**
ReCOVery	0.511±0.002	0.458±0.004	0.617±0.104	0.544±0.128	0.510±0.001	0.439±0.001	0.511±0.004	0.426±0.007	0.514±0.001	0.443±0.004	0.751±0.009	0.726±0.009	0.878±0.014	0.888±0.013	**0.937±0.021**	**0.929±0.017**
MC Fake	0.471±0.003	0.474±0.005	0.523±0.002	0.452±0.004	0.519±0.005	0.434±0.003	0.487±0.001	0.474±0.005	0.501±0.002	0.453±0.005	0.485±0.103	0.461±0.010	0.762±0.015	0.778±0.014	**0.886±0.013**	**0.833±0.013**
LIAR	0.480±0.006	0.382±0.005	0.494±0.029	0.414±0.030	0.475±0.002	0.417±0.006	0.524±0.002	0.490±0.004	0.542±0.002	0.507±0.004	0.482±0.013	0.500±0.013	0.575±0.016	0.572±0.015	**0.675±0.020**	**0.672±0.019**
PAN2020	0.508±0.005	0.337±0.004	0.498±0.032	0.389±0.079	0.526±0.003	0.467±0.009	0.508±0.005	0.512±0.004	0.523±0.006	0.489±0.009	0.745±0.014	0.724±0.014	0.732±0.020	0.723±0.021	**0.774±0.014**	**0.769±0.017**

Table 4: Detection performance on five datasets (best in red, second-best in blue).

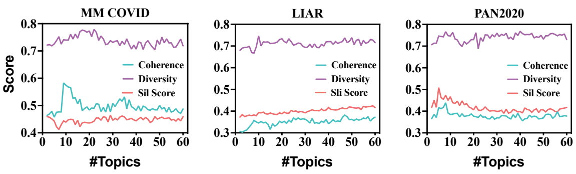


Figure 3: Coherence, Diversity and Sil Score with different number of topics on three datasets.

4.2 Topic Modeling Validation

Topic modeling is pivotal to constructing the $\mathcal{HG}$. In this section, we specifically validate the choices for the optimal topic numbers and their impact on the detection performance.

Optimal Topic Number. We use a multi-metric approach to select the optimal number of topics for each dataset, considering topic coherence for interpretability, topic diversity for variety, and the Silhouette Score for topic separation and compactness. The evaluation spans a range of topic numbers, from 3 to 60. Ideally, the optimal number of topics corresponds to the point where all three metrics reach their peak values, but as depicted in Figs. 3 and 8 no point meets this criterion. Therefore, we compromise by selecting six topic numbers for each dataset, which yield the highest or near-highest values for at least one metric.

The Impact of Topic Numbers on the Detection Performance. As depicted in Fig. 10, we observe slight variations in the performance of LEG4FD across different topic numbers on each dataset, while the optimal topic numbers for each dataset are: 44 for MM COVID, 58 for ReCOVery, 8 for MC Fake, 10 for LIAR, and 40 for PAN2020.

4.3 Ablation Study

In this ablation study, we assess the impact of each model components by omitting them one at a time: '$\oslash\mathcal{HG}$' excludes the heterogeneous graph, relying only on LLM-extracted news embeddings for detection; '$\oslash$T' and '$\oslash$E' remove topic and entity nodes from the graph, respectively; and '$\oslash$CR' omits the consistency learning module.

From the results in Table 5, we observe a notable decrement in performance when directly use LMM-extracted embeddings for fake news detection, exemplified by the case of '$\oslash\mathcal{HG}$'. After incorporating the heterogeneous graph into the training process, as demonstrated by '$\oslash$E', '$\oslash$T', and '$\oslash$CR', the results are enhanced across all datasets. Such performance gaps before and after engaging with $\mathcal{HG}$ further support our motivation to learn high-level semantics for fake news detection. Meanwhile, the better performance of '$\oslash$E' and '$\oslash$T', compared to '$\oslash\mathcal{HG}$', showcase that each of them benefits our model from capturing the nuances of fake news. As proposed to engage unlabeled news for a fine-gained training of the detector, the consistency loss is capable of improving the overall performance around 2% on the five datasets, by comparing '$\oslash$CR' and LEG4FD.

Datasets	Methods	Acc	Pre	Rec	F1
MM COVID	LEG4FD $\oslash\mathcal{HG}$	0.634±0.053	0.539±0.216	0.555±0.074	0.481±0.130
	LEG4FD $\oslash$ E	0.924±0.021	0.928±0.020	0.919±0.021	0.920±0.021
	LEG4FD $\oslash$ T	0.938±0.020	0.937±0.022	0.942±0.019	0.939±0.020
	LEG4FD $\oslash$ CR	0.950±0.019	0.950±0.018	0.948±0.020	0.948±0.020
	LEG4FD	**0.974±0.010**	**0.975±0.010**	**0.973±0.009**	**0.973±0.010**
ReCOVery	LEG4FD $\oslash\mathcal{HG}$	0.685±0.005	0.526±0.217	0.504±0.006	0.418±0.015
	LEG4FD $\oslash$ E	0.870±0.017	0.864±0.016	0.865±0.020	0.854±0.019
	LEG4FD $\oslash$ T	0.884±0.015	0.870±0.016	0.880±0.019	0.870±0.017
	LEG4FD $\oslash$ CR	0.904±0.020	0.910±0.027	0.908±0.019	0.891±0.023
	LEG4FD	**0.938±0.020**	**0.930±0.018**	**0.937±0.021**	**0.929±0.017**
MC Fake	LEG4FD $\oslash\mathcal{HG}$	0.818±0.007	0.414±0.009	0.501±0.004	0.453±0.006
	LEG4FD $\oslash$ E	0.839±0.013	0.761±0.015	0.800±0.015	0.754±0.016
	LEG4FD $\oslash$ T	0.854±0.011	0.781±0.009	0.829±0.011	0.798±0.012
	LEG4FD $\oslash$ CR	0.869±0.009	0.809±0.009	0.842±0.013	0.818±0.014
	LEG4FD	**0.894±0.012**	**0.826±0.015**	**0.886±0.013**	**0.833±0.013**
LIAR	LEG4FD $\oslash\mathcal{HG}$	0.556±0.021	0.534±0.123	0.523±0.026	0.443±0.066
	LEG4FD $\oslash$ E	0.626±0.027	0.649±0.040	0.629±0.027	0.625±0.027
	LEG4FD $\oslash$ T	0.638±0.024	0.670±0.061	0.636±0.027	0.633±0.028
	LEG4FD $\oslash$ CR	0.654±0.029	0.671±0.035	0.653±0.027	0.650±0.031
	LEG4FD	**0.678±0.021**	**0.765±0.019**	**0.675±0.020**	**0.672±0.019**
PAN2020	LEG4FD $\oslash\mathcal{HG}$	0.558±0.073	0.515±0.165	0.557±0.071	0.496±0.125
	LEG4FD $\oslash$ E	0.718±0.069	0.767±0.067	0.711±0.076	0.704±0.087
	LEG4FD $\oslash$ T	0.731±0.049	0.770±0.050	0.728±0.050	0.724±0.052
	LEG4FD $\oslash$ CR	0.7571±0.025	0.766±0.025	0.757±0.023	0.755±0.024
	LEG4FD	**0.771±0.017**	**0.798±0.019**	**0.774±0.014**	**0.769±0.017**

Table 5: Ablation results.

4.4 Further Analysis

Scales of Feature Propagation. The scales of feature propagation determine the local and global semantics to be explored. In LEG4FD, we control the scales using two parameters s_l and s_g, as presented in Sec. 3.4. We vary their values and depict their influence in Figs. 6 and 9. It is evident that the model performs best when s_l is around 5 denoting that the local semantics within 5-hops is optimal,

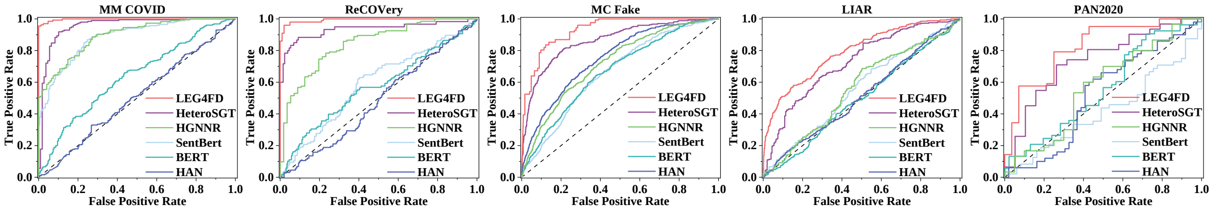


Figure 4: ROC curves on five datasets.

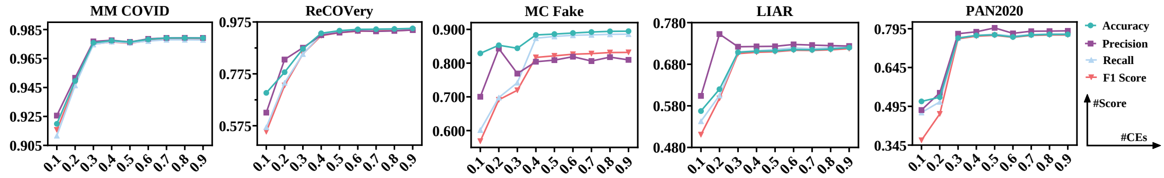


Figure 5: Sensitivity to λ_{ce}.

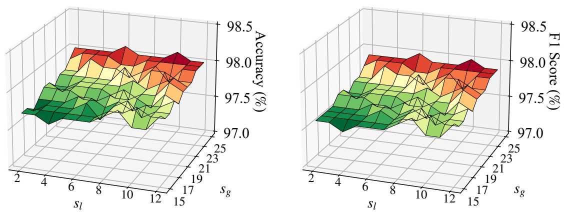


Figure 6: Sensitivity to s_l and s_g on MM COVID w.r.t. accuracy and F1.

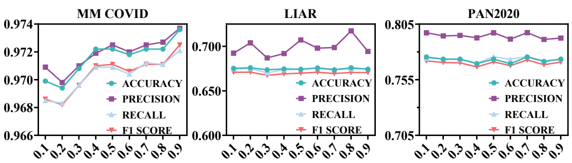


Figure 7: Sensitivity to λ_g on three datasets.

while a larger s_g always lead to better performance since more global information are involves.

Impact of λ_{ce} and λ_{cr}. Those two hyperparameters balance the weights of training loss on labeled and unlabeled news. A higher value of λ_{cr} makes the model to place more emphasis on unlabeled data, whereas a larger λ_{ce} will stress more on leveraging supervision from labeled news. The results in Fig. 11 suggest that increasing λ_{cr} is beneficial to the detection performance when it is below 0.6, but stressing more will detoriate the performance. In contrast, from Fig. 5, we see that increasing the proportion of loss from labeled news constantly improves the detection performance.

Impact of λ_g. In contrast, from Fig. 7, we observe that, for the majority of datasets, our model maintains steady performance despite variations in the weights of global semantics.

Method	Source of Features		Semantics	
	Time (s/epoch)	Mem (MB)	Time (s/epoch)	Mem (MB)
TextCNN	0.115	649.413	1.951	816.292
TextGCN	0.066	538.879	0.343	1354.532
HAN	9.976	1908.109	43.643	2528.107
BERT	0.11	958.879	0.803	3040.097
SentenceBERT	0.131	962.392	2.102	2626.038
HGNNR4FD	1.078	988.765	2.956	2098.223
HeteroSGT	0.238	547.826	0.98	2302.512
LEG4FD	0.056	740.312	0.0678	2043.563

Table 6: Running time & GPU memory cost.

Computational Costs. In additional to its superior performance, we also highlight LEG4FD's efficiency, showcasing reduced time per training epoch and lower overall GPU memory usage, as detailed in Table 6.

5 Conclusion

In this paper, we propose a novel method, LEG4FD, to take the advantage of LLMs for detecting fake news. We first employ LLM as the enhancer to extract news, entities, news and their corresponding features using a set of potent prompts. By modeling the extracted data as a heterogeneous graph, we then propose an effective feature propagation algorithm to encode both the local and global semantics which simultaneous involves the training signal from unlabeled news to enrich the training of the detector. Through extensive experiments on five widely-used datasets, we showcase the new state-of-the-art in fake news detection.

Limitations. In this work, we only explore ChatGPT and API provided by OpenAI for enhancing fake news detection. Extending our method to work with other open-sourced LLMs and tunning LLM particularly for fake news detection are important directions for future efforts.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI*, volume 34, pages 549–556.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *ACL*, pages 627–638.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *SIGIR*, pages 2051–2055.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *AAAI*, volume 35, pages 81–89.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Xiaoxin He, Xavier Bresson, Thomas Laurent, and Bryan Hooi. 2023. Explanations as features: Llm-based features for text-attributed graphs. *arXiv preprint arXiv:2305.19523*.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *ICWSM*, volume 11, pages 759–766.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *ACL/IJCNLP*, pages 754–763.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *WWW*, pages 2915–2921.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2023. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP-IJCNLP*, pages 4821–4830.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *ACL*.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *ACL*, pages 1363–1380.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *WWW*, pages 1148–1158.

Maxime Prieur, Souhir Gahbiche, Guillaume Gadek, Sylvain Gatepaille, Kilian Vasnier, and Valerian Justine. 2023. K-pop and fake facts: from texts to smart alerting for maritime security. In *ACL*, pages 510–517.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Francisco Rangel, Anastasia Giachanou, Bilal Hisham Hasan Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR workshop proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *EMNLP*, pages 3347–3363.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *KDD*, pages 395–405.

Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. 2023. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *WSDM*, pages 51–59.

Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. 2018. Fake news detection: a deep learning approach. *SMU Data Science Review*, 1(3):10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *KDD*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Lingwei Wei, Dou Hu, Yantong Lai, Wei Zhou, and Songlin Hu. 2022. A unified propagation forest-based framework for fake news detection. In *COLING*, pages 2769–2779.

Zhengxuan Wu and Desmond C Ong. 2021. Context-guided bert for targeted aspect-based sentiment analysis. In *AAAI*, volume 35, pages 14094–14102.

Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, Shan Xue, and Hao Fan. 2023. Heterogeneous graph neural network via knowledge relations for fake news detection. In *SSDM*, pages 1–11.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *WWW*, pages 2501–2510.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *KDD*, KDD '22, page 2253–2262, New York, NY, USA. Association for Computing Machinery.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT*, pages 1480–1489.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*, volume 33, pages 7370–7377.

Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *WWW*, pages 3465–3476.

Yuchen Zhang, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2024. Heterogeneous subgraph transformer for fake news detection. In *WWW*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *CIKM*, pages 3205–3212.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Experimental Details

A.1 Baselines

For a fair evaluation of the overall detection performance and considering the availability of additional sources, we compared LEG4FD with seven representative baseline algorithms including:

- **textCNN** (Kim, 2014) is designed to capture localized patterns and features within input texts. It utilizes Convolutional Neural Network layers (CNNs) to small windows of words in the text to extract patterns and features for news classification.

- **textGCN** (Yao et al., 2019) represents input texts as nodes in a graph, employing graph convolutional operations on both the textual content of each document and the graph structure. This process aims to learn effective representations for fake news detection.

- **HAN** (Yang et al., 2016), or Hierarchical Attention Network, employs attention mechanisms to represent intricate relationships at both word-sentence and sentence-article levels, enhancing its ability to capture hierarchical features for improved fake news detection performance.

- **BERT** (Devlin et al., 2018) is a prominent transformer-based language model. In our experimentation, we utilize the embedded representation of the [CLS] token from BERT for the task of fake news classification.

- **SentenceBERT** (Reimers and Gurevych, 2019) is an extension of BERT that is specifically designed for sentence embeddings. It uses siamese and triplet network structures during training to generate semantically meaningful sentence embeddings

- **HGNNR4FD** (Xie et al., 2023) models news articles in a heterogeneous graph and incorporates external entity knowledge from Knowledge Graphs to enhance the learning of news representations for fake news detection.

- **HeteroSGT** (Zhang et al., 2024) proposes a heterogeneous subgraph transformer to exploit subgraphs in the news heterogeneous graph that contains relations between news articles, topics, and entities.

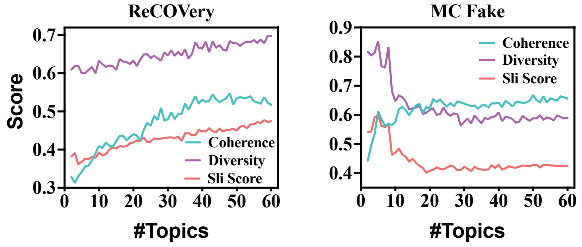
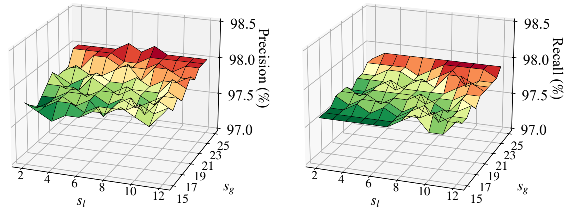


Figure 8: Coherence, Diversity and Sil Score with different number of topics on ReCOVery and MC Fake.



(a) (b)

Figure 9: Sensitivity to s_l and s_g on MM COVID w.r.t. precision and recall.

A.2 Hyperparameter and Computational Settings

Hyperparameters. For constructing $\mathcal{HG}$, we choose the optimal number of topics $|\mathbb{T}|$ for each dataset through the comprehensive topic model evaluation detailed in Sec. 4.2. We perform a grid search to determine the remaining hyperparameters, with the search space defined as follows:

Feature propagation scale s^l: [2, 12]
Feature propagation scale s^g: [15, 25]
Trade-off parameter λ_g: [0.1, 0.9]
Cross-entropy loss weight λ_{ce}: [0.1, 0.9]
Consistency loss weight λ_{cr}: [0.1, 1.0]

Computational Environment. All the experiments are conducted on a Rocky Linux 8.6 (Green Obsidian) server with 12-core CPU and 1 NVIDIA Volta GPU (with 30G RAM).

A.3 Sensitivity to s_l and s_g

In addition to Fig 6 in Sec. 4.2, we can see that our model performs best with $s_l = 5$ and $s_g = 25$.

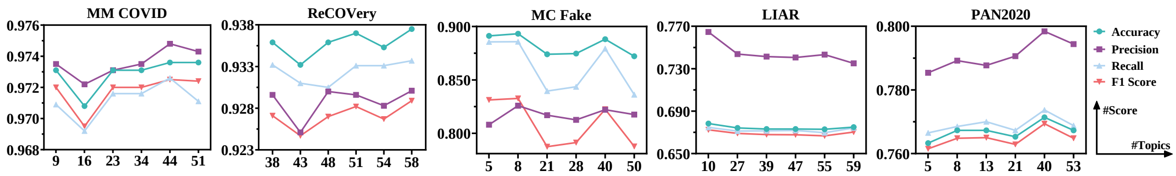


Figure 10: LEG4FD's performance on datasets with different numbers of topics.

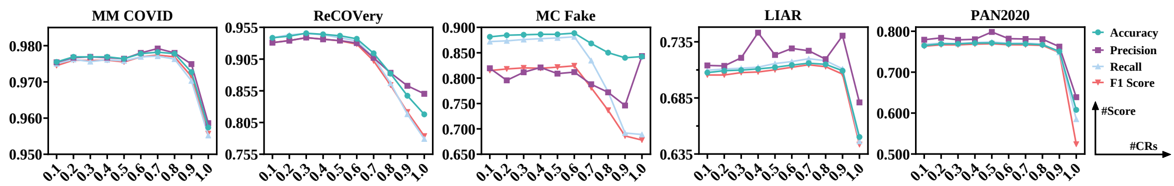


Figure 11: Sensitivity to λ_{cr}.