On Fake News Detection with LLM Enhanced Semantics Mining

Xiaoxiao Ma[1,2, †*], Yuchen Zhang[1,3†], Kaize Ding[4], Jian Yang[1], Jia Wu[1], Hao Fan[3]

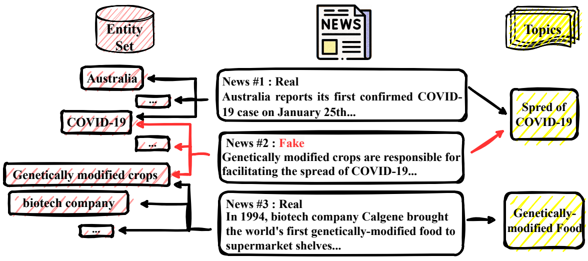
[1]School of Computing, Macquarie University, Sydney, Australia
[2]Amazon Machine Learning, Sydney, Australia
[3]School of Information Management, Wuhan University, Hubei, China
[4]Department of Statistics and Data Science, Northwestern University, IL, USA
{xiaoxiao.ma2@hdr, yuchen.zhang3@hdr, jian.yang@, jia.wu@}mq.edu.au
{kaize.ding@northwestern.edu} {hfan@whu.edu.cn}

Abstract

Large language models (LLMs) have emerged as valuable tools for enhancing textual features in various text-related tasks. Despite their superiority in capturing the lexical semantics between tokens for text analysis, our preliminary study on two popular LLMs, i.e., GPT-3.5 and Llama2, shows that simply applying news embeddings from LLMs is ineffective for fake news detection. Such embeddings only encapsulate the language styles between tokens. Meanwhile, the high-level semantics among named entities and topics, which reveal the deviating patterns of fake news, have been ignored. Therefore, we propose a topic model together with a set of specially designed prompts to extract topics and real entities from LLMs and model the relations among news, entities, and topics as a heterogeneous graph to facilitate investigating news semantics. We then propose a Generalized Page-Rank model and a consistent learning criterion for mining the local and global semantics centered on each news piece through the adaptive propagation of features across the graph. Our model shows superior performance on five benchmark datasets over seven baseline methods and the efficacy of the key ingredients has been thoroughly validated.

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; Ma et al., 2024). To alleviate the far-reaching consequences, many fake news detection methods probe the information dissemination process or social structure (Mehta et al., 2022; Ma et al., 2021) 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.



News #1 : Real
Australia reports its first confirmed COVID-19 case on January 25th...

News #2 : Fake
Genetically modified crops are responsible for facilitating the spread of COVID-19...

News #3 : Real
In 1994, biotech company Calgene brought the world's first genetically-modified food to supermarket shelves...

Figure 1: Irregular co-occurrence of meaningful entities in fake news on a specific topic (red arrows).

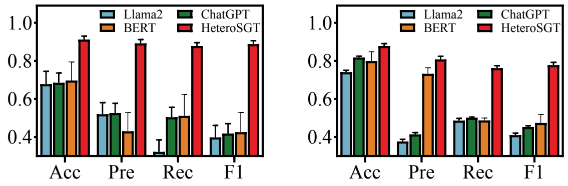
Following the latter approach, in which news embeddings are critical for providing a discriminatory description of authentic and fake news, we are propelled to enhance them with Large Language Models (LLMs), which have been renowned for their remarkable capabilities in language understanding, and context modeling (Thota et al., 2018; Zhao et al., 2023; Li et al., 2024b). A fundamental question that guides our research in this under-explored realm is, *“Are the LLMs output news embeddings effective for fake news detection?"*

To answer this question, we conducted a preliminary study by comparing the detection performance of an MLP classifier trained using news embeddings extracted from GPT-3.5[1], Llama2[2], BERT (Kenton and Toutanova, 2019) and HeteroSGT (Zhang et al., 2024), respectively. From the results depicted in Fig. 2 (and Table 8), we

[*]. Work done at Macquarie University.
[†]. Contributed Equally

[1]. https://api.openai.com
[2]. https://llama.meta.com

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 508–521
November 12-16, 2024 ©2024 Association for Computational Linguistics



(a) ReCOVery (b) MC Fake

Figure 2: A comparison between fake news detection performance on two datasets w.r.t. **acc**uracy, **pre**cision, **rec**all and **F1** score.

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	✗	✓	✗	✓	✓	✗
LESS4FD (Ours)	✗	✓	✗	✓	✓	**CR**

Table 1: Overview of existing methods. Comparisons are made upon the source information, the semantics each method explores, and how they enforce learning on unlabeled data.

found that simply applying the LLMs and BERT extracted news embeddings is ineffective for fake news detection because they primarily focus on lexical semantics between tokens. When fake news mimics the language styles of authentic news, this approach fails.

On the other hand, the better performance of a recent method, HeteroSGT, which investigates the high-level semantic relations among news, entities, and topics for fake news detection, affirms previous findings that the knowledge of real entities and topics is crucial for identifying fake news (Huang et al., 2019; Xie et al., 2023; Jeong et al., 2022). Taken news #2 depicted in Fig. 1 as an example, it is fake because the named entity 'Genetically modified crops' is not 'responsible' for 'COVID-19' when discussing the '#Spread of COVID-19'. These discoveries signify high-level semantics for fake news detection, however, two further sub-problems exist:

***P1**. How can we apply LLMs to explore high-level news semantics?* From the above study, we affirm that the exploration of high-level semantics enables the model to acquire a better perception of deeper contextual nuances, which encompass fabricated knowledge among entities with real meaning on a particular topic (Zhang et al., 2024), for distinguishing fake news. We identify the keys for high-level semantics exploration using LLMs are to extract *meaningful entities* and *topics*.

***P2.** How can we identify the irregular semantics in fake news?* Given the LLM-derived entities and topics, one can aggregate their features to enhance the centered news embeddings for fake news detection. But this primarily focuses on the information within individual news pieces (local semantics), lacking the ability to explicitly explore the broader range of knowledge across news pieces (global semantics) to identify narrative inconsistencies and manipulations in fake news. For example, in detecting news #2 as fake, we identify the relation between 'COVID-19' and 'Genetically modified crops' to be irregular because they rarely co-appear in other news discussions about the '#Spread of COVID-19'. Therefore, to identify the deviating semantic patterns of fake news, it is crucial to investigate both the local semantics of individual articles and the global semantics across news pieces.

To address ***P1***, by prompting LLMs for entity extraction, we first propose a refined topic model that summarizes news topics through LLM-generated embeddings. We then construct a heterogeneous graph to model the relationships among news, entities, and topics by representing them as nodes and connecting them with edges, which facilitates further exploration of local and global news semantics.

For ***P2***, we apply short- and long-scale feature propagation centered on news nodes to encapsulate the local and global semantics into news representations. With these two scales of feature propagation, we can identify inconsistencies between each individual news text and the broader knowledge across news, and involve unlabeled news for training with our specially designed consistency training criterion. Our major contributions are:

- Our preliminary study uncovers two fundamental problems that should be addressed to incorporate LLMs for advancing the detection of fake news;
- We introduce an LLM-enhanced topic model and devise potent prompts for querying LLMs. Our proposed method, LESS4FD, not only captures local semantics surrounding individual news and the global semantics spanning across the dataset to identify the inconsistencies of fake news but also allows a flexible consistency regularization on unlabeled data for refining the news representation;
- Extensive experiments on five real-world datasets demonstrate the superiority of our method over seven baseline methods and confirm our design choices.

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 (Ding et al., 2022, 2024), such as word-word relations (Yao et al., 2019; Linmei et al., 2019; Li et al., 2023), news dissemination graphs (Ma et al., 2018, 2023; 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; 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 compare our work and the existing methods in Table 1.

2.2 LLMs for Feature Mining

LLMs such as GPT (Brown et al., 2020), Llama2 (Touvron et al., 2023), and pre-trained language models like BERT (Kenton and Toutanova, 2019) have emerged as powerful tools for feature mining due to their remarkable adaptability in language understanding and sentiment analysis (Min et al., 2023; Liu et al., 2023; Wu and Ong, 2021). LLMs for feature mining primarily focus on enriching the embeddings of texts. The most straightforward application involves feeding the output features into specific models for tasks such as time series analysis and graph learning (Jin et al., 2023).

To get more specific information and further enrich the textual features, more advanced methods prompt LLMs to generate supplementary content, such as related knowledge and background information (Min et al., 2023). This additional content is then combined with the original texts for downstream modeling (He et al., 2023; Li et al., 2024a). In summary, LLMs showcase their potential for advancing various natural language processing-related tasks, and this paper addresses the two prior recognized sub-problems to take advantage of LLMs for fake news detection.

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 la-

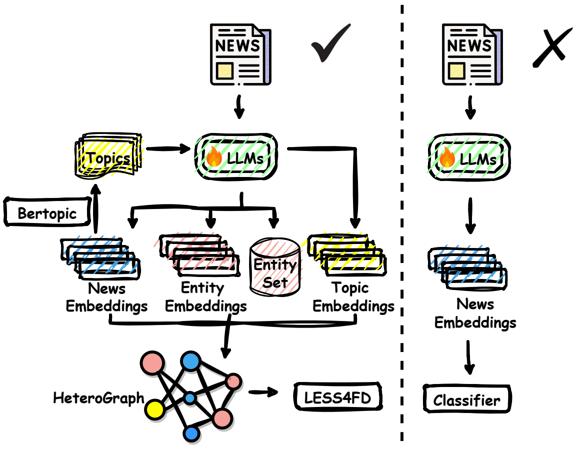


Figure 3: Heterogeneous graph construction.

bels 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. 2, we further investigate LLMs, particularly GPT-3.5 and Llama2, to address our devised ***P1*** as follows. For brevity, we use LLM to denote GPT-3.5 or Llama2.

Entity Extraction. For news entity extraction, we prompt the LLM following Table 2 for identifying specific entities in all news pieces including persons, dates, locations, organizations, and miscellaneous entities[3].

News and Entity Embedding. We obtain the news embeddings and entity embeddings by directly querying the API provided by OpenAI[2] and Meta[3] to encode the corresponding lexical semantics in the text. The resulting news embeddings are processed as $\mathbf{X}^n$, and the entity embeddings are stored in $\mathbf{X}^e$.

Topic Modeling. In addition to entities, modeling the topics across news pieces not only enables us to summarize the news focus and link different news pieces, but also to explore the relation between the target news and entities in another news,

```
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 ... >
GPT-3.5:
"PERSON": ["Elon Musk", ... ],    "DATE": ["Dec 2030", ... ],
"LOC": ["Mars", ... ],    "ORG": ["SpaceX", ... ],
"MISC": ["CEO", ... ]
```

Table 2: Prompt for entity extraction.

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:

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

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.

For replication purposes, we detail the practical settings in entity extraction, embedding, and topic modeling in Sec. 4, accompanied by an 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,

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

entities, and topics, denoted by $\mathbb{X}$. This approach addresses our recognized ***P1*** 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. 2, 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-layer 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 are 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 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.

3.6 Training Objective and Fake News Detection

Combing both the supervised loss and consistency loss, the overall training objective of LESS4FD (**L**LM **E**nhanced **S**emantic**S** mining for fake news detection) can be formulated as:

$$\arg\min_{\boldsymbol{w},\boldsymbol{\theta}} \lambda_{ce} \mathcal{L}_{sup} + (1 - \lambda_{ce}) \mathcal{L}_{con}, \tag{7}$$

where λ_{ce} 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 Appendix A.1.

Baselines. We compare LESS4FD[6] against seven representative baselines in text classification and fake news detection, including **textCNN** (Kim, 2014), **textGCN** (Yao et al., 2019), **BERT** (Kenton and Toutanova, 2019), **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.3.

Experimental Settings. To test the overall performance, we adopt the two most popular LLMs, GPT-3.5 and Llama2, to extract entities, topics, and news embeddings.

We perform 10-fold cross-validation (using a split ratio of 80%-10%-10% for training, validation, and test) and report the averaged results along with the standard deviations regarding five mostly-used metrics: Accuracy (Acc), macro-precision (Pre), macro-recall (Rec), macro-F1 (F1), and the AUC-ROC curve. We conduct all case studies with GPT-3.5 because of its better performance, and for brevity, we refer to the implementation using GPT-3.5 as 'LESS4FD*' and the implementation with Llama2 as 'LESS4FD$^\diamond$'. Detailed hyperparameter settings are provided in Appendix A.4.

4.1 Fake New Detection Performance

Overall Performance. The results summarized in Tables 3, and 4, and Fig. 5 reveal that our method surpasses all baseline models w.r.t. the five evaluation metrics. The performance gaps, which are over 5% on MM COVID and 2% on the rest datasets, affirm the effectiveness of our approach in investigating the LLM-enhanced news semantics for fake news detection. It is also worth noting that there are firm differences between LESS4FD* and LESS4FD$^\diamond$, which indicate both GPT-3.5- and Llama2-derived embeddings are effective. By comparison with different categories of baselines, we also observe that:

High-level Semantic Exploration is Pivotal. Despite the effectiveness of traditional classifiers like TextCNN, TextGCN, HAN, BERT, and Sentence-

[4] https://www.politifact.com
[5] https://www.gossipcop.com
[6] https://github.com/XiaoxiaoMa-MQ/Less4FD

Model	MM COVID		ReCOVery		MC Fake		LIAR		PAN2020	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
TextCNN	0.564±0.038	0.492±0.104	0.649±0.002	0.458±0.004	0.816±0.004	0.474±0.005	0.556±0.002	0.382±0.005	0.503±0.002	0.337±0.004
TextGCN	0.691±0.160	0.642±0.245	0.733±0.004	0.544±0.128	0.697±0.142	0.452±0.004	0.487±0.039	0.414±0.030	0.495±0.032	0.389±0.079
HAN	0.829±0.009	0.838±0.009	0.694±0.003	0.439±0.001	0.834±0.004	0.434±0.003	0.559±0.003	0.417±0.006	0.494±0.005	0.467±0.009
BERT	0.744±0.110	0.711±0.103	0.697±0.003	0.426±0.007	0.799±0.005	0.474±0.005	0.522±0.004	0.490±0.004	0.519±0.005	0.512±0.004
SentenceBert	0.761±0.004	0.729±0.006	0.687±0.006	0.443±0.004	0.828±0.002	0.453±0.005	0.566±0.002	0.507±0.004	0.524±0.005	0.489±0.009
HGNNR4FD	0.732±0.017	0.755±0.021	0.783±0.008	0.726±0.009	0.818±0.010	0.461±0.010	0.544±0.013	0.500±0.013	0.690±0.014	0.724±0.014
HeteroSGT	0.924±0.011	0.916±0.012	0.912±0.010	0.888±0.013	0.878±0.012	0.778±0.014	0.582±0.017	0.572±0.015	0.720±0.021	0.723±0.021
LESS4FD$^\diamond$	0.973±0.011*	0.972±0.011*	0.917±0.017*	0.897±0.020*	0.883±0.006*	0.787±0.008*	**0.689±0.034***	0.658±0.035*	0.731±0.037*	0.727±0.037*
LESS4FD*	**0.974±0.010***	**0.973±0.010***	**0.938±0.020***	**0.929±0.017***	**0.894±0.012***	**0.833±0.013***	0.678±0.021*	**0.672±0.019***	**0.771±0.017***	**0.769±0.017***

Table 3: Detection performance w.r.t accuracy and F1 score on five datasets (best in **red**, second-best in blue). * indicates that the performance improvement is statistically significant at a 95% confidence level ($\alpha = 0.05$) compared to the best baseline results.

Model	MM COVID		ReCOVery		MC Fake		LIAR		PAN2020	
	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec	Pre	Rec
TextCNN	0.484±0.173	0.560±0.004	0.449±0.107	0.511±0.002	0.530±0.159	0.471±0.003	0.447±0.185	0.480±0.006	0.309±0.119	0.508±0.005
TextGCN	0.716±0.240	0.694±0.181	0.697±0.183	0.617±0.104	0.524±0.173	0.523±0.002	0.493±0.047	0.494±0.029	0.392±0.144	0.498±0.032
HAN	0.836±0.007	0.834±0.004	0.435±0.201	0.510±0.001	0.444±0.103	0.519±0.005	0.501±0.005	0.475±0.002	0.457±0.135	0.526±0.003
BERT	0.705±0.010	0.723±0.112	0.430±0.214	0.511±0.004	0.732±0.003	0.487±0.001	0.522±0.002	0.524±0.002	0.541±0.005	0.508±0.005
SentenceBert	0.786±0.002	0.730±0.006	0.645±0.167	0.514±0.001	0.464±0.006	0.501±0.002	0.565±0.002	0.542±0.002	0.508±0.009	0.523±0.006
HGNNR4FD	0.882±0.016	0.648±0.021	0.771±0.006	0.751±0.009	0.456±0.010	0.485±0.103	0.559±0.009	0.482±0.013	0.677±0.014	0.745±0.014
HeteroSGT	0.918±0.012	0.912±0.012	0.892±0.014	0.878±0.014	0.808±0.012	0.762±0.015	0.579±0.016	0.575±0.016	0.731±0.021	0.732±0.020
LESS4FD$^\diamond$	0.972±0.011*	0.972±0.010*	0.905±0.017*	0.894±0.022*	0.811±0.014*	0.806±0.014*	0.728±0.046*	**0.712±0.034***	0.777±0.030*	0.749±0.037*
LESS4FD*	**0.975±0.010***	**0.973±0.009***	**0.930±0.018***	**0.937±0.021***	**0.826±0.015***	**0.886±0.013***	**0.765±0.019***	0.675±0.020*	**0.798±0.019***	**0.774±0.014***

Table 4: Detection performance w.r.t precision and recall on five datasets (best in **red**, second-best in blue).

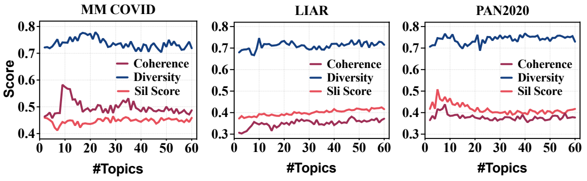


Figure 4: Coherence, Diversity, and Sil Score with the different numbers of topics on three datasets.

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 and analyzing the relations and features of news, entities, and topics.

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 has deteriorated due to the insufficient exploration of global and local semantics. Specifically, HGNNR4FD only focuses on local semantics, while HeteroSGT suffers from information loss through random walks. Our method addresses these issues by mining global and local semantics at lower computational costs (see Table 6).

Overall, we attribute LESS4FD's superiority to the investigation of high-level semantics in news text and mining global and local semantics in $\mathcal{HG}$, which have been further validated in Sec. 4.3.

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. 4 and 10 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. 8, we observe slight variations in the performance of LESS4FD 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 component by omitting them one at a time: '$\oslash\mathcal{HG}$' excludes the heterogeneous graph, relying only on LLM-extracted news embeddings for detec-

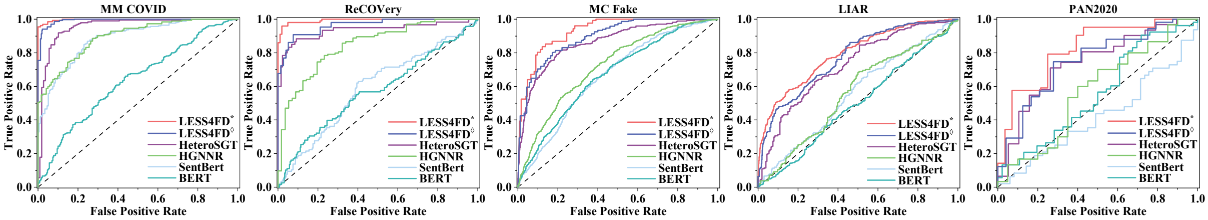
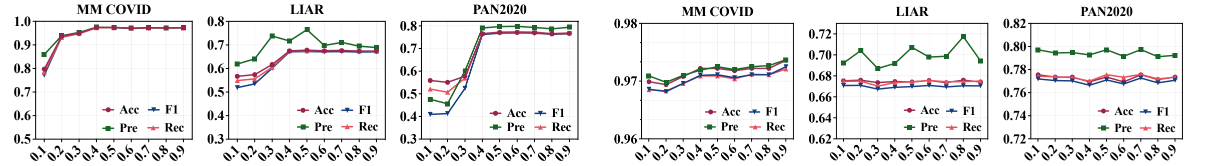


Figure 5: ROC curves on five datasets.



(a) λ_{ce}

(b) λ_g

Figure 6: Sensitivity to λ_{ce} and λ_g on three datasets.

Datasets	Methods	Acc	Pre	Rec	F1
MM COVID	LESS4FD* ⊘ $\mathcal{HG}$	0.634±0.053	0.539±0.216	0.555±0.074	0.481±0.130
	LESS4FD* ⊘ E	0.924±0.021	0.928±0.020	0.919±0.021	0.920±0.021
	LESS4FD* ⊘ T	0.938±0.020	0.937±0.022	0.942±0.019	0.939±0.020
	LESS4FD* ⊘ CR	0.950±0.019	0.950±0.018	0.948±0.020	0.948±0.020
	LESS4FD*	**0.974±0.010**	**0.975±0.010**	**0.973±0.009**	**0.973±0.010**
LIAR	LESS4FD* ⊘ $\mathcal{HG}$	0.556±0.021	0.534±0.123	0.523±0.026	0.443±0.066
	LESS4FD* ⊘ E	0.626±0.027	0.649±0.040	0.629±0.027	0.625±0.027
	LESS4FD* ⊘ T	0.638±0.024	0.670±0.061	0.636±0.027	0.633±0.028
	LESS4FD* ⊘ CR	0.654±0.029	0.671±0.035	0.653±0.027	0.650±0.031
	LESS4FD*	**0.678±0.021**	**0.765±0.019**	**0.675±0.020**	**0.672±0.019**
ReCOVery	LESS4FD*⊘$\mathcal{HG}$	0.685±0.052	0.526±0.051	0.504±0.053	0.418±0.053
	LESS4FD*⊘E	0.870±0.017	0.864±0.016	0.865±0.020	0.854±0.019
	LESS4FD*⊘T	0.884±0.015	0.870±0.016	0.880±0.019	0.870±0.017
	LESS4FD*⊘CR	0.904±0.020	0.910±0.027	0.908±0.019	0.891±0.023
	LESS4FD*	**0.938±0.020**	**0.930±0.018**	**0.937±0.021**	**0.929±0.017**
MC Fake	LESS4FD*⊘$\mathcal{HG}$	0.818±0.007	0.414±0.009	0.501±0.004	0.453±0.006
	LESS4FD*⊘E	0.839±0.013	0.761±0.015	0.800±0.015	0.754±0.016
	LESS4FD*⊘T	0.854±0.011	0.781±0.009	0.829±0.011	0.798±0.012
	LESS4FD*⊘CR	0.869±0.009	0.809±0.009	0.842±0.013	0.818±0.014
	LESS4FD*	**0.894±0.012**	**0.826±0.015**	**0.886±0.013**	**0.833±0.013**
PAN2020	LESS4FD*⊘$\mathcal{HG}$	0.558±0.073	0.515±0.165	0.557±0.071	0.496±0.125
	LESS4FD*⊘E	0.718±0.069	0.767±0.067	0.711±0.076	0.704±0.087
	LESS4FD*⊘T	0.731±0.049	0.770±0.050	0.728±0.050	0.724±0.052
	LESS4FD*⊘CR	0.7571±0.025	0.766±0.025	0.757±0.023	0.755±0.024
	LESS4FD*	**0.771±0.017**	**0.798±0.019**	**0.774±0.014**	**0.769±0.017**

Table 5: Ablation results of LESS4FD* on five datasets.

tion; '⊘T' and '⊘E' remove topic and entity nodes from the graph, respectively; and '⊘CR' omits the consistency learning module.

From the results in Tables 5 and 10, we observe a notable decrement in performance when directly using LLM-extracted embeddings for fake news detection, exemplified by the case of '⊘$\mathcal{HG}$'. After incorporating the heterogeneous graph into the training process, as demonstrated by '⊘E', '⊘T', and '⊘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 '⊘E' and '⊘T', compared to '⊘$\mathcal{HG}$', 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 '⊘CR' and LESS4FD.

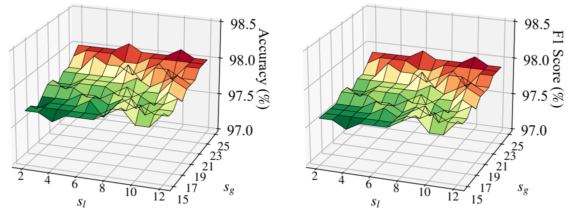


Figure 7: Sensitivity to s_l and s_g on MM COVID w.r.t. accuracy and F1 score.

4.4 Further Analysis

We further study the impacts of different parameter settings and training cost of our news representation learning method. We use LESS4FD* unless specified.

Scales of Feature Propagation. The scales of feature propagation determine the local and global semantics to be explored. Both scales can be adjusted upon two parameters s_l and s_g, as presented in Sec. 3.4. We vary their values and depict their influence in Figs. 7 and 11. It is evident that the model performs best when s_l is around 5 denoting that the local semantics within 5-hops is optimal, while a larger s_g always leads to better performance since more global information is involved.

Impact of λ_{ce}. This hyperparameter balances the weights of training loss on labeled and unlabeled news. A higher value of λ_{ce} makes the model emphasize more on labeled data. To assess its impact, we adjust λ_{ce} between 0.1 and 0.9 and depict the results in Fig. 6(a). We see that increasing λ_{ce} is beneficial to the detection performance, particularly when it remains below 0.4. Beyond this point, marginal fluctuations in performance emerge across datasets and the optimal range for λ_{ce} consistently lies between 0.4 and 0.6.

Impact of λ_g. λ_g is to regularize the training signal from the exploration of global semantics. As illustrated in Fig. 6(b), we find that our model maintains almost steady performance despite variations in the weights of global semantics.

Impact of Potential Data Contamination. At the time of this study, all datasets had already been published before the LLMs' training date and they might have been involved in tuning the textual tokens in LLMs. However, for our task of fake news detection, we clarify that such potential data contamination merely impacts our research findings because: **1)** The LLMs we use, specifically GPT-3.5 and Llama2, are primarily trained for text-generation rather than fake news detection; **2)** In our preliminary experiments, as reported in Fig. 2 and Table 8, the news embeddings derived from these LLMs proved to be ineffective for fake news detection; and **3)** Through our extensive ablation study, we demonstrate that our performance gains stem from the novel model design of exploring high-level semantics as well as the local and global information, which is typically ignored in the tokenized training text of LLMs.

To validate this claim, we further compare the performance of our method with that of the best baseline method, HeteroSGT, by incorporating entities, topics, and news embeddings derived from GPT-3.5 into both models. As both our method and HeteroSGT utilize the same sets of entities, topics, news, and embeddings, this setup allows for a fair comparison of the model designs for fake news detection. According to the results presented in Table 9, our design consistently demonstrates superior detection performance.

Computational Costs. In addition to the detection performance improvement, we also evaluate LESS4FD's efficiency, showcasing reduced time per training epoch with moderate GPU memory usage, as detailed in Table 6.

5 Conclusion

In this paper, we propose LESS4FD to take advantage of LLMs for enhancing semantics mining for fake news detection. We first employ LLMs as the enhancers to extract news, entities, topics, 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 into news embeddings to enrich the training of the detector. Through extensive experiments on five widely-used datasets, our method demonstrates better performance than seven baseline methods while the efficacy of key ingredients is further validated in the case studies.

Method	MM COVID		MC Fake	
	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	0.110	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	0.980	2302.512
LESS4FD*	0.056	740.312	0.068	2043.563
LESS4FD$^\diamond$	0.067	878.235	0.082	2371.381

Table 6: Running time & GPU memory cost.

Limitations. In this work, we only adopt the two most popular LLMs as enhancers to explore the news semantics. Extending our method to tuning LLMs, particularly for fake news detection is an important direction for future efforts.

Ethical issues. The datasets utilized in our research for detecting fake news are widely accessed and publicly available for academic research. Our proposed method exclusively relies on the textual content of news articles from these datasets as input, without requiring any additional user-specific information (e.g., personal identifiers) or user social information (e.g., retweet/comment behavior). We employed publicly accessible APIs provided by OpenAI and Meta to obtain embeddings. Our prompts, which are made publicly available, are used exclusively for extracting entities and topics from LLMs. Therefore, our method ensures minimal risk of privacy infringement.

Applications. Detecting fake news is critical due to its significant implications for society, politics, and individual decision-making. Our proposed model demonstrates efficacy in distinguishing authentic and false content, which could contribute to mitigate the spread of false information and public distrust.

Acknowledgements

This work was supported by the Australian Research Council Projects LP210301259 and DP230100899, and Macquarie University Data Horizons Research Centre.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*.

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*.

Kaize Ding, Xiaoxiao Ma, Yixin Liu, and Shirui Pan. 2024. Divide and denoise: Empowering simple models for robust semi-supervised node classification against label noise. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Kaize Ding, Jianling Wang, James Caverlee, and Huan Liu. 2022. Meta propagation networks for graph few-shot semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*.

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

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, Adam Perold, Yann LeCun, and Bryan Hooi. 2023. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *ICLR*.

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*.

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*.

Qi Huang, Chuan Zhou, Jia Wu, Mingwen Wang, and Bin Wang. 2019. Deep structure learning for rumor detection on twitter. In *IJCNN*.

Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2022. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*.

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*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

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

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

Shiyang Li, Jianshu Chen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, et al. 2024a. Explanations from large language models make small reasoners better. In *2nd Workshop on Sustainable AI*.

Yichuan Li, Kaize Ding, and Kyumin Lee. 2023. Grenade: Graph-centric language model for self-supervised representation learning on text-attributed graphs. *arXiv preprint arXiv:2310.15109*.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024b. A survey of graph meets large language model: Progress and future directions. In *IJCAI*.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

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

Xiaoxiao Ma, Ruikun Li, Fanzhen Liu, Kaize Ding, Jian Yang, and Jia Wu. 2024. Graph anomaly detection with few labels: A data-centric approach. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. 2021. A comprehensive survey on graph anomaly detection with deep learning. *IEEE TKDE*.

Xiaoxiao Ma, Jia Wu, Jian Yang, and Quan Z Sheng. 2023. Towards graph-level anomaly detection via deep evolutionary mapping. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

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*.

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*.

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*.

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*.

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*.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *KDD*.

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*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *ACL*.

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*.

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

Zhengxuan Wu and Desmond C Ong. 2021. Context-guided bert for targeted aspect-based sentiment analysis. In *AAAI*.

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*.

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

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *KDD*.

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

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*.

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

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

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*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*.

A Experimental Details

A.1 Datasets

The statistical details of the five datasets are summarized in Table 7.

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 7: Statistics of datasets.

A.2 Preliminary Experiment Results

Our preliminary experiment results with Llama2, ChatGPT, BERT and HeteroSGT on ReCOVery and MC Fake datasets were summarized in Table 8.

A.3 Baselines

For a fair evaluation of the overall detection performance and considering the availability of additional sources, we compared LESS4FD 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 (Kenton and Toutanova, 2019) 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.

A.4 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. For a fair comparison between LESS4FD* and LESS4FD$^\diamond$, we use the same set of entities, topics, and their embeddings from GPT-3.5, while the news embeddings are derived from GPT-3.5 and Llama2, respectively. 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]

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

A.5 Addition Experimental Results

Optimal Topic Number. We depict the Coherence, Diversity, and Silhouette Score with different numbers of topics on ReCOVery and MC Fake in Fig. 10 and similar to that on MM COVID, LIAR, and PAN2020, no point meets the criterion where all three metrics reach their peak values.

Fake News Detection Performance. From Tables 4 and 3, we see that our proposed method LESS4FD performs better than all baseline methods. To demonstrate the statistical significance of performance improvement, we conduct further pairwise t-test at a 95% confidence level ($a = 0.05$). The results in Tables 11, 12, 13, and 14 show that the performance improvement is significant.

Ablation Study. In addition to the ablation study on LESS4FD*, we report the results on LESS4FD$^\diamond$ in Table 10. Similar to that in Table 5, we can see

Method	ReCOVery				MC Fake			
	Acc	Pre	Rre	F1	Acc	Pre	Rec	F1
Llama2	0.678 ± 0.067	0.520 ± 0.061	0.322 ± 0.063	0.398 ± 0.063	0.741 ± 0.010	0.377 ± 0.011	0.486 ± 0.012	0.410 ± 0.011
GPT-3.5	0.685 ± 0.052	0.526 ± 0.051	0.504 ± 0.053	0.418 ± 0.053	0.818 ± 0.007	0.414 ± 0.009	0.501 ± 0.004	0.453 ± 0.006
BERT	0.697 ± 0.003	0.430 ± 0.214	0.511 ± 0.004	0.426 ± 0.007	0.799 ± 0.005	0.732 ± 0.003	0.487 ± 0.001	0.474 ± 0.005
HeteroSGT	0.912 ± 0.018	0.892 ± 0.020	0.878 ± 0.018	0.888 ± 0.018	0.878 ± 0.013	0.808 ± 0.016	0.762 ± 0.013	0.778 ± 0.014

Table 8: Preliminary experiment results.

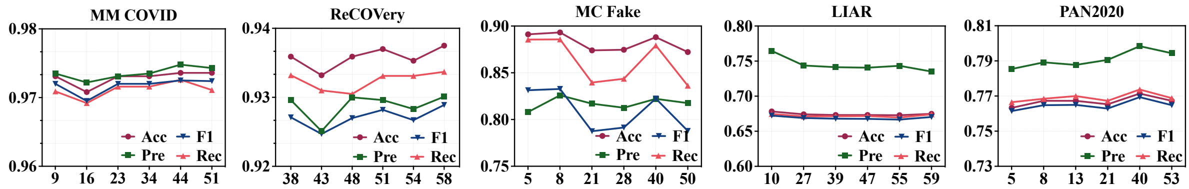


Figure 8: Performance of LESS4FD* on datasets with different numbers of topics.

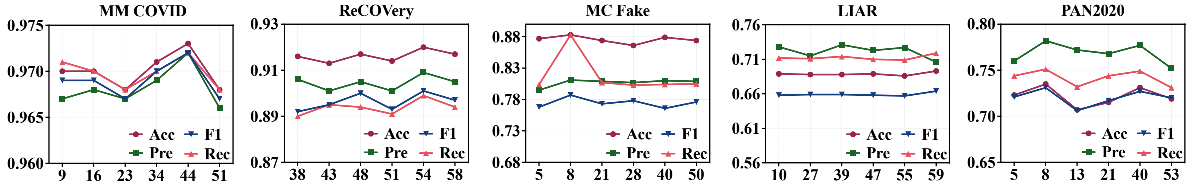


Figure 9: Performance of LESS4FD$^{\diamond}$ on datasets with different numbers of topics.

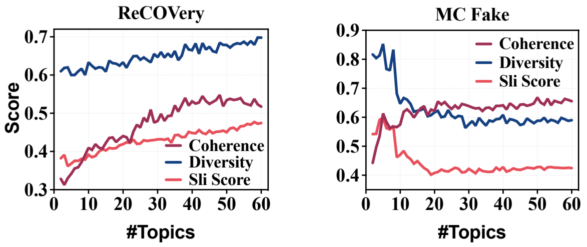


Figure 10: Coherence, Diversity and Sil Score with different numbers of topics on ReCOVery and MC Fake.

Datasets	Methods	Acc	Pre	Rec	F1
MM COVID	HeterSGT(GPT-3.5)	0.949±0.011	0.939±0.012	0.955±0.010	0.946±0.013
	LESS4FD*	0.974±0.010	0.975±0.010	0.973±0.009	0.973±0.010
LIAR	HeterSGT(GPT-3.5)	0.644±0.013	0.640±0.015	0.638±0.015	0.638±0.016
	LESS4FD*	0.678±0.021	0.765±0.019	0.675±0.020	0.672±0.019
PAN2020	HeterSGT(GPT-3.5)	0734±0.020	0.735±0.021	0.726±0.019	0.727±0.020
	LESS4FD*	0.771±0.017	0.798±0.019	0.774±0.014	0.769±0.017

Table 9: Comparison with HeteroSGT's performance using LLM-derived entities, topics, and embeddings.

that the key ingredients consistently yield better detection performance using Llama2 and GPT-3.5.

A.6 Sensitivity to s_l and s_g

In addition to Fig. 7 in Sec. 4.2, we can see that our model performs best with $s_l = 5$ and $s_g = 25$ w.r.t. precision and recall on MM COVID.

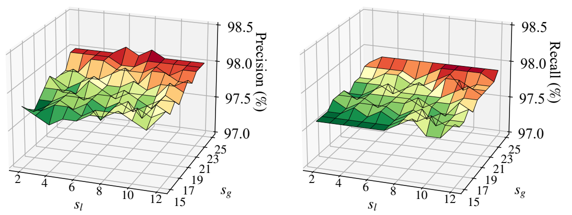


Figure 11: Sensitivity to s_l and s_g on MM COVID w.r.t. precision and recall.

Datasets	Methods	Acc	Pre	Rec	F1
MM COVID	LESS4FD$^{\diamond}$ $\oslash \mathcal{HG}$	0.612±0.018	0.592±0.020	0.578±0.018	0.518±0.018
	LESS4FD$^{\diamond}$ $\oslash E$	0.923±0.019	0.921±0.020	0.922±0.019	0.921±0.020
	LESS4FD$^{\diamond}$ $\oslash T$	0.941±0.019	0.938±0.022	0.941±0.022	0.937±0.021
	LESS4FD$^{\diamond}$ $\oslash CL$	0.943±0.018	0.944±0.019	0.942±0.018	0.941±0.019
	LESS4FD$^{\diamond}$	**0.973±0.011**	**0.972±0.011**	**0.972±0.010**	**0.972±0.011**
ReCOVery	LESS4FD$^{\diamond}$ $\oslash \mathcal{HG}$	0.678±0.067	0.520±0.061	0.322±0.063	0.398±0.063
	LESS4FD$^{\diamond}$ $\oslash E$	0.814±0.020	0.793±0.026	0.7705±0.019	0.779±0.022
	LESS4FD$^{\diamond}$ $\oslash T$	0.852±0.021	0.876±0.025	0.824±0.021	0.822±0.023
	LESS4FD$^{\diamond}$ $\oslash CL$	0.887±0.020	0.890±0.025	0.841±0.021	0.839±0.023
	LESS4FD$^{\diamond}$	**0.917±0.017**	**0.905±0.017**	**0.894±0.022**	**0.897±0.020**
MC Fake	LESS4FD$^{\diamond}$ $\oslash \mathcal{HG}$	0.741±0.010	0.377±0.011	0.486±0.012	0.410±0.011
	LESS4FD$^{\diamond}$ $\oslash E$	0.794±0.011	0.706±0.012	0.776±0.013	0.743±0.010
	LESS4FD$^{\diamond}$ $\oslash T$	0.820±0.011	0.713±0.058	0.796±0.012	0.760±0.014
	LESS4FD$^{\diamond}$ $\oslash CL$	0.834±0.008	0.745±0.057	0.798±0.009	0.767±0.011
	LESS4FD$^{\diamond}$	**0.883±0.006**	**0.811±0.014**	**0.806±0.014**	**0.787±0.008**
LIAR	LESS4FD$^{\diamond}$ $\oslash \mathcal{HG}$	0.521±0.023	0.563±0.062	0.478±0.022	0.393±0.023
	LESS4FD$^{\diamond}$ $\oslash E$	0.613±0.021	0.671±0.056	0.604±0.027	0.609±0.029
	LESS4FD$^{\diamond}$ $\oslash T$	0.629±0.024	0.692±0.032	0.624±0.032	0.619±0.032
	LESS4FD$^{\diamond}$ $\oslash CL$	0.658±0.021	0.656±0.044	0.654±0.025	0.647±0.025
	LESS4FD$^{\diamond}$	**0.689±0.034**	**0.728±0.046**	**0.712±0.034**	**0.658±0.035**
PAN2020	LESS4FD$^{\diamond}$ $\oslash \mathcal{HG}$	0.528±0.062	0.511±0.088	0.573±0.065	0.447±0.095
	LESS4FD$^{\diamond}$ $\oslash E$	0.694±0.055	0.684±0.051	0.622±0.047	0.683±0.055
	LESS4FD$^{\diamond}$ $\oslash T$	0.706±0.053	0.703±0.040	0.700±0.047	0.698±0.054
	LESS4FD$^{\diamond}$ $\oslash CL$	0.729±0.050	0.740±0.044	0.729±0.051	0.721±0.053
	LESS4FD$^{\diamond}$	**0.731±0.037**	**0.777±0.030**	**0.749±0.037**	**0.727±0.037**

Table 10: Ablation results of LESS4FD$^{\diamond}$ on five datasets.

Dataset	A-TextCNN	A-TextGCN	A-HAN	A-BERT	A-SentenceBert	A-HGNNR4FD	A-HeteroSGT
MM COVID	1.4E-17	8.0E-07	1.1E-17	2.0E-04	3.5E-11	3.8E-09	4.8E-10
ReCOVery	6.3E-10	9.6E-17	1.4E-08	3.1E-08	4.8E-08	2.7E-14	4.0E-05
MC Fake	2.2E-16	8.2E-05	1.2E-12	3.4E-17	5.1E-16	9.8E-14	7.2E-05
LIAR	5.0E-12	1.6E-10	7.1E-12	1.1E-13	1.8E-11	1.5E-12	4.1E-09
PAN2020	5.8E-13	1.3E-11	4.0E-14	2.6E-13	3.9E-13	4.4E-05	9.7E-04
Dataset	**B-TextCNN**	**B-TextGCN**	**B-HAN**	**B-BERT**	**B-SentenceBert**	**B-HGNNR4FD**	**B-HeteroSGT**
MM COVID	1.1E-17	1.0E-06	3.2E-09	2.6E-04	2.9E-13	8.8E-10	7.1E-11
ReCOVery	2.7E-17	1.3E-14	4.5E-16	7.9E-16	6.7E-16	1.3E-12	8.1E-05
MC Fake	2.1E-16	3.6E-05	1.5E-13	3.4E-17	6.5E-16	2.0E-14	1.1E-06
LIAR	1.4E-15	4.5E-11	2.1E-15	2.6E-17	5.1E-15	2.4E-15	1.4E-10
PAN2020	4.8E-14	1.6E-15	6.5E-14	3.1E-16	3.9E-17	1.1E-08	4.6E-04

Table 11: Pairwise t-test on Accuracy. A-TextCNN denotes the t-test results between Less4FD$^{\diamond}$ and baseline methods, while B-TextCNN denotes the t-test results between Less4FD* and baselines.

Dataset	A-TextCNN	A-TextGCN	A-HAN	A-BERT	A-SentenceBert	A-HGNNR4FD	A-HeteroSGT
MM COVID	1.9E-10	3.0E-04	3.9E-10	6.5E-14	7.0E-15	4.3E-13	7.5E-11
ReCOVery	1.1E-10	4.1E-07	1.6E-09	2.1E-06	1.5E-11	1.3E-18	2.4E-03
MC Fake	1.1E-04	2.9E-09	2.4E-11	1.4E-16	2.0E-12	1.8E-11	1.1E-04
LIAR	3.4E-04	3.7E-09	3.4E-10	2.8E-08	4.1E-07	1.2E-05	1.9E-04
PAN2020	3.1E-13	4.0E-11	1.1E-07	8.7E-14	4.2E-08	1.0E-03	1.4E-04
Dataset	**B-TextCNN**	**B-TextGCN**	**B-HAN**	**B-BERT**	**B-SentenceBert**	**B-HGNNR4FD**	**B-HeteroSGT**
MM COVID	1.8E-10	2.8E-04	2.9E-15	2.1E-15	3.1E-16	2.1E-10	3.2E-08
ReCOVery	4.8E-11	1.4E-07	7.4E-10	1.0E-06	5.4E-12	1.0E-17	1.3E-04
MC Fake	4.5E-05	1.1E-09	7.9E-12	2.3E-19	1.5E-14	3.1E-13	3.9E-05
LIAR	2.5E-05	1.3E-13	1.9E-12	3.4E-11	4.1E-10	3.9E-18	5.2E-13
PAN2020	1.4E-14	2.1E-12	7.9E-09	1.4E-12	8.1E-13	5.6E-12	1.8E-07

Table 12: Pairwise t-test on Precision. A-TextCNN denotes the t-test results between Less4FD$^{\diamond}$ and baseline methods, while B-TextCNN denotes the t-test results between Less4FD* and baselines.

Dataset	A-TextCNN	A-TextGCN	A-HAN	A-BERT	A-SentenceBert	A-HGNNR4FD	A-HeteroSGT
MM COVID	4.4E-13	4.3E-04	6.2E-10	1.0E-04	3.3E-14	1.2E-10	8.3E-10
ReCOVery	8.9E-11	1.8E-09	2.1E-13	2.3E-14	1.3E-15	8.6E-17	1.8E-05
MC Fake	2.6E-15	2.6E-14	1.2E-15	1.9E-17	2.7E-15	1.6E-08	8.6E-06
LIAR	8.3E-09	8.2E-09	2.3E-12	5.8E-13	1.0E-08	4.9E-18	3.6E-10
PAN2020	3.9E-10	2.7E-15	2.3E-16	5.4E-18	6.1E-08	6.1E-05	1.5E-04
Dataset	**B-TextCNN**	**B-TextGCN**	**B-HAN**	**B-BERT**	**B-SentenceBert**	**B-HGNNR4FD**	**B-HeteroSGT**
MM COVID	8.3E-16	4.7E-04	3.6E-17	1.2E-04	6.8E-12	8.2E-10	1.4E-08
ReCOVery	8.7E-13	4.8E-10	2.0E-11	2.3E-12	1.9E-13	7.8E-16	7.4E-05
MC Fake	7.2E-15	5.3E-14	4.1E-15	1.4E-16	8.1E-15	6.2E-10	2.7E-12
LIAR	4.8E-11	3.7E-11	5.1E-08	2.6E-13	5.0E-10	1.5E-15	2.7E-07
PAN2020	1.8E-10	4.3E-16	3.0E-17	1.3E-18	1.4E-12	7.3E-09	4.0E-04

Table 13: Pairwise t-test on Recall. A-TextCNN denotes the t-test results between Less4FD$^{\diamond}$ and baseline methods, while B-TextCNN denotes the t-test results between Less4FD* and baselines.

Dataset	A-TextCNN	A-TextGCN	A-HAN	A-BERT	A-SentenceBert	A-HGNNR4FD	A-HeteroSGT
MM COVID	2.1E-12	1.5E-05	1.7E-13	1.6E-09	4.6E-09	3.3E-12	2.1E-12
ReCOVery	5.1E-15	5.6E-10	1.4E-11	4.6E-12	5.1E-11	2.2E-13	9.6E-05
MC Fake	8.8E-10	1.2E-11	3.3E-16	1.2E-13	7.0E-17	3.1E-15	1.4E-04
LIAR	3.2E-08	1.4E-13	1.4E-16	7.3E-14	4.5E-13	4.5E-13	2.1E-08
PAN2020	1.8E-14	8.3E-13	1.3E-08	2.2E-11	5.6E-18	6.8E-05	5.9E-05
Dataset	**B-TextCNN**	**B-TextGCN**	**B-HAN**	**B-BERT**	**B-SentenceBert**	**B-HGNNR4FD**	**B-HeteroSGT**
MM COVID	2.3E-12	1.5E-05	9.2E-18	1.7E-09	2.5E-12	1.2E-17	7.2E-10
ReCOVery	2.6E-15	1.7E-10	5.3E-12	1.9E-12	1.4E-10	1.9E-11	1.4E-05
MC Fake	9.9E-11	1.4E-12	3.9E-17	1.2E-17	8.0E-14	3.2E-16	9.0E-12
LIAR	9.3E-09	1.3E-15	5.5E-13	1.6E-10	8.3E-11	1.2E-14	2.8E-13
PAN2020	1.2E-15	5.2E-14	9.1E-10	1.9E-13	2.0E-13	7.2E-09	1.0E-08

Table 14: Pairwise t-test on F1 score. A-TextCNN denotes the t-test results between Less4FD$^{\diamond}$ and baseline methods, while B-TextCNN denotes the t-test results between Less4FD* and baselines.