Collaborative Evolution: Multi-Round Learning Between Large and Small Language Models for Emergent Fake News Detection

Ziyi Zhou, Xiaoming Zhang*, Shenghan Tan, Litian Zhang, Chaozhuo Li

Beihang University,
{ziyizhou, yolixs, tsh21371459, litianzhang}@buaa.edu.cn, lichaozhuo1991@gmail.com

Abstract

The proliferation of fake news on social media platforms has exerted a substantial influence on society, leading to discernible impacts and deleterious consequences. Conventional deep learning methodologies employing small language models (SLMs) suffer from the necessity for extensive supervised training and the challenge of adapting to rapidly evolving circumstances. Large language models (LLMs), despite their robust zero-shot capabilities, have fallen short in effectively identifying fake news due to a lack of pertinent demonstrations and the dynamic nature of knowledge. In this paper, a novel framework Multi-Round Collaboration Detection (MRCD) is proposed to address these aforementioned limitations. The MRCD framework is capable of enjoying the merits from both LLMs and SLMs by integrating their generalization abilities and specialized functionalities, respectively. Our approach features a two-stage retrieval module that selects relevant and up-to-date demonstrations and knowledge, enhancing in-context learning for better detection of emerging news events. We further design a multi-round learning framework to ensure more reliable detection results. Our framework MRCD achieves SOTA results on two real-world datasets Pheme and Twitter16, with accuracy improvements of 7.4% and 12.8% compared to using only SLMs, which effectively addresses the limitations of current models and improves the detection of emergent fake news detection.

Introduction

The proliferation of fake news on social networks has caused significant impact and harm to society (Zhou and Zafarani 2020). In order to achieve automated detection of fake news, numerous deep learning-based approaches have been proposed (Zhang et al. 2024c; Wu and Hooi 2023; Hu et al. 2021). The traditional methods primarily utilize small language models (SLMs) like BERT (Devlin et al. 2018) to extract features from news content or propagation paths in social networks to accomplish the classification task, demonstrating promising performance across multiple fake news detection datasets.

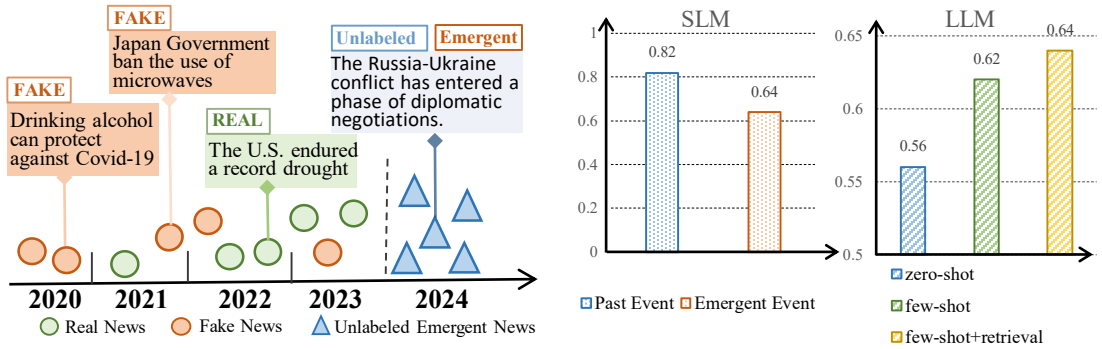
Despite the satisfactory performance of current methods on specific datasets, these methods are mainly based on supervised training, whereas in reality, a large number of emergent news events occur frequently (Olan et al. 2022). Manual annotation of data is both time-consuming and costly, leading to insufficient data for supervised training of models (Yin et al. 2024; Gangireddy et al. 2020). Moreover, the continuous evolution of news results in differences between the distribution of emergent events and manually annotated data as Figure 1(a) illustrates. These factors collectively lead to traditional SLMs being unable to adapt well to emergent events, thus failing to demonstrate satisfactory performance in practical applications as Figure 1(b) presented.

Large Language Models (LLMs) have shown remarkable abilities on various NLP applications due to their robust zero-shot capabilities (Gruver et al. 2024; Kojima et al. 2022; Zhang et al. 2024a). However, experiments in Figure 1(b) have shown that LLMs do not exhibit strong detection capabilities for fake news in zero-shot and few-shot scenarios. This might be attributed to two main reasons: Firstly, while LLMs demonstrate strong task adaptation abilities, the potential of LLMs for fake news detection remains inactive due to the lack of suitable demonstrations in zero-shot environments. Secondly, news events are continually evolving while frozen-parameters LLMs lack external dynamic knowledge to assist in judging emerging events.

Due to the limitations of LLMs in fake news detection, recent researches (Su, Cardie, and Nakov 2023; Wan et al. 2024; Chen and Shu 2023) have opted to utilize LLMs to provide additional assistance to SLMs for more accurate detection. Dell (Wan et al. 2024) utilizes LLMs as agents to form a social network by generating user comments and employ LLMs to extract additional information such as sentiment and knowledge to aid SLMs in fake news detection. ARG leverages LLMs to analyze news content and provide rationales to help SLMs in detection (Hu et al. 2024). Recent studies also leverage LLMs to extract more effective external knowledge to assist in assessing the authenticity of news (Liu et al. 2024). Although the utilization of LLMs have shown improved detection capabilities, these methods still rely on a substantial amount of data for training the SLMs as they only use LLMs as an auxiliary to SLMs. Furthermore, they neglect the fact that LLMs possess strong learning and generalization capabilities for detecting fake news.

To address the dependency on data in current methods and the issue of current models' inability to accurately detect emergent events, we propose a novel **M**ulti-**R**ound

*Xiaoming Zhang is the corresponding author.



(a) The distribution shift and lack of annotation are two challenges for emergent fake news detection.

(b) SLM's performance on past and emergent events, and LLM's performance on emergent events.

Figure 1: (a) illustrates two major challenges for emergent fake news detection, the distribution shift and lack of annotations. (b) firstly demonstrates that SLMs perform well on testing past events after training, but its judgment capability significantly decreases on emergent events. It then shows that LLM performs poorly on emergent events directly in zero-shot scenarios. However, its detection effectiveness improves after using annotated data for few-shot learning and retrieval-augmented methods. The two experiments are conducted on Twitter16 dataset with RoBERTa as the SLM and Llama3-8B as the LLM.

Collaboration **D**etection framework, dubbed **MRCD** that combines the generalization capability of LLMs with the specialized expertise of SLMs to achieve more accurate detection of the authenticity of emerging news events. Inspired by the success of in-context learning and retrieval-augmented generation across numerous downstream tasks (Chen et al. 2023; Sun et al. 2023b; Lewis et al. 2020), we also employ a novel two-stage retrieval module to select better demonstrations and acquire the latest knowledge relevant to the emergent news events. Since there are no similar manually labeled news datasets available for emergent events, we use a news corpus and online search engines in the first-stage retrieval. This ensures we fetch the latest news articles closest to the news under detection. The extracted news content is then assigned with pseudo labels to serve as demonstrations for the LLM's in-context learning. The second-stage retrieval utilizes wikipedia as external knowledge corpus to retrieve latest knowledge relevant to the emergent events to provide the LLM more common sense information. After performing in-context learning with the LLM and pre-trained SLM, we use both models to inference the unlabeled news articles and devise a data selection method to divide them into clean data pool and noisy data pool. Finally, we design a multi-round learning framework where the LLM and SLM collaborate to obtain a more generalized SLM and transform all data into samples with clean labels. Our framework MRCD achieves SOTA results on two real-world datasets Pheme and Twitter16, with accuracy improvements of 7.4% and 12.8% compared to using only SLMs.

Our contributions are summarized as follows:

- We introduce collaborative efforts between LLMs and SLMs for fake news detection and further devise a multi-round sample selection process to enhance detection accuracy in unsupervised emergent news events settings.
- To enable the LLM to have a better semantic understanding of emergent news events and to access real-time knowledge, a two-stage retrieval module is proposed to select better demonstrations by utilizing online search engines and unlabeled news corpus while acquiring the latest and accurate knowledge from Wikipedia.
- Extensive experiments on real-world datasets demonstrate our method MRCD achieved SOTA performance and improves SLM's performance significantly by collaborative multi-round learning between the LLM and the SLM.

Related Work

Fake News Detection

Fake News Detection typically adopts a binary classification framework, distinguishing between real and fake news articles. Recent researches can be mainly divided into three categories: content-based methods, knowledge-augmented methods and propagation-based methods. Content-based methods primarily discern the authenticity of news by extracting distinct semantic information from real and fake news (Khattar et al. 2019; Chen et al. 2022). Knowledge-augmented methods utilize external knowledge corpus such as wikidata to provide common sense knowledge to aid model's detection (Hu et al. 2021; Zhang et al. 2024c; Sun et al. 2023a). Propagation-based methods focuses on employing the propagation paths of news to identify misinformation (Shu et al. 2020; Zhang et al. 2024b,d).

Due to the remarkable capabilities of LLMs across various NLP tasks, recent research has emerged attempting to apply LLMs to fake news detection. Dell (Wan et al. 2024) utilizes LLMs as agents to form a social network by generating user comments and employ LLMs to extract additional information such as sentiment and knowledge to aid SLMs in fake news detection. ARG leverages LLMs to analyze news content and provide rationales to help SLMs in detection (Hu et al. 2024). Recent studies also leverage LLMs to extract effective external knowledge to assist in assessing the authenticity of news (Liu et al. 2024). Although these methods have shown promising performance, they have two main shortcomings. Firstly, they still require a large amount of data to train the SLMs and lack good generalization ability

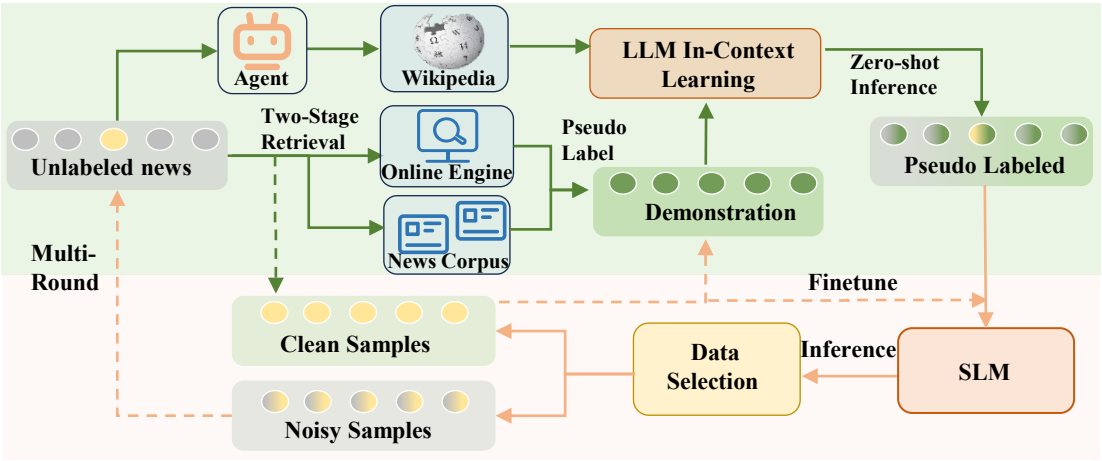


Figure 2: The architecture of MRCD. The $\rightarrow$ denotes the first round of learning while the $\dashrightarrow$ denotes the process for the subsequent rounds of learning.

for emerging news events (Silva et al. 2021; Zhou et al. 2024; Zhang et al. 2023). Secondly, they only utilize LLMs to provide additional knowledge to the small model, ignoring the large model's inherent ability for task reasoning and learning.

In-Context Learning

Recent research has demonstrated that LLMs can exhibit excellent performance in numerous zero-shot and few-shot downstream tasks through in-context learning (Li and Qiu 2023; Sun et al. 2023b; Chen et al. 2023; Ye et al. 2023). Among these researches, a considerable of them has focused on selecting better demonstrations (Liu et al. 2022; Wu et al. 2023). Unlike most studies that concentrate on utilizing algorithms like BM25 to select relevant data from annotated training datasets as demonstrations for test samples, Lyu and Min (Min et al. 2022; Lyu et al. 2023) shows that in-context learning benefits mainly from the correct distribution of the inputs and the labels. Inspired by this, we choose to select demonstrations from both search engines and news corpus and assign pseudo labels for LLMs' in-context learning. This allows us to obtain examples closely related to the emergent news events under detection, avoiding significant distribution discrepancies between demonstrations and test data.

Retrieval-Augmented LLMs

Despite the outstanding performance of LLMs across various tasks, they still face challenges such as hallucination, outdated knowledge and high fine-tuning costs (Gao et al. 2023; Li et al. 2017). Retrieval-Augmented methods have shown promising effects in many knowledge-intensive tasks as they allow for continuous knowledge updates and integration of domain-specific information (Lewis et al. 2020), such as Question Answering (Khattab et al. 2022), dialog generation (Lin et al. 2023) and commonsense reasoning (Cheng et al. 2023). To help LLM detect the authenticity of emergent news events, we employed a two-stage retrieval process to retrieve demonstrations and real-time knowledge for LLM.

Problem Formulation

This paper primarily aims to address the detection of emergent news events. Let ε denote a set of news events. Each event e is associated with a timestamp indicating its occurrence. We sort the events based on these timestamps, arranging them from earliest to latest : $\varepsilon = (e_1, t_1), (e_2, t_2), ..., (e_n, t_n)$. Then the events are divided into two past events and future events by their timestamps. The past events $\{X_e^s, Y_e^s\} = \{x_{e,i}^s, y_{e,i}^s\}_{i=1}^K$ are labeled posts and the data pertaining to future events $\{X_e^t\} = \{x_{e,i}^t\}_{i=K+1}^N$ lacks any form of annotation. The data of past events are used to initialize a SLM. The objective is for the LLM $\mathcal{L}$ and the initialized SLM $\mathcal{S}$ to collaborate effectively in detecting the authenticity of emergent news events.

Methodology

In this section, we introduce our proposed framework MRCD which investigates fostering collaboration between LLM and SLM. It incorporates a two-stage retrieval module to retrieve unlabeled news articles as demonstrations and utilizes knowledge from wikipedia to enhance the judgment capability of the LLM. The LLM then makes zero-shot predictions for the emergent news events under detection. Simultaneously, the pre-trained SLM directly infers pseudo-labels for the news. A data selection module is then designed to divide the data into clean and noisy subsets, with the noisy data undergoing another iteration for detection and the clean data using as demonstrations for the LLM and training samples for the SLM. The overview of MRCD is displayed in Figure 2.

Two-Stage Retrieval Module

Demonstration Retrieval Traditional demonstration selection often involves finding data related to the samples under detection from annotated training datasets for few-shot in-context learning (Li and Qiu 2023; Ye et al. 2023; Wu et al. 2023; Zhang et al. 2022). However, due to significant distribution differences between continuously updated news

events and pre-annotated training data, directly using annotated training data as examples may not enable the LLM to perform well in in-context learning. Recent research has also found that the success of in-context learning depends more on the semantic consistency between the test data and the demonstration data, rather than the correctness of the labels (Liu et al. 2022; Min et al. 2022; Lyu et al. 2023). Therefore, we utilize online search engines[1] $\mathcal{W}$ to retrieve the latest news, avoiding mismatches between data in static text corpus and emergent news events. To avoid the singularity of data selection from search engines, we additionally choose a news corpus (Przybyla 2020; Zhao et al. 2021) $\mathcal{C}$ to supplement and ensure diversity in the retrieved data.

Let x denotes the news content, $N_w = \{w_1, w_2, ..., w_n\}$ denotes the retrieved news from search engines $\mathcal{W}$ and $N_c = \{c_1, c_2, ...c_m\}$ denotes retrieved news from news corpus $\mathcal{C}$. We utilize BM25 algorithm to extract the most semantically and structurally similar top-k data from N_w and N_c as demonstrations $N_k = \{d_1, d_2, ..., d_k\}$ as Equation 1:

$$N_k = \text{BM25}_{\text{top-k}}(x, N_w \cup N_c) \tag{1}$$

Directly incorporating unlabeled news diminishes the LLM's ability to discern authentic news from fake, as the LLM relies heavily on labeled demonstrations to learn task-specific nuances (Liu et al. 2022). Additionally, simply assigning pseudo labels such as "real" or "fake" can result in copy effects, where the model mimics the surface characteristics of the labels rather than understanding the underlying truthfulness of the content (Lyu et al. 2023). Inspired by (Lyu et al. 2023; Li et al. 2018), we use semantic synonyms as pseudo labels. Each unlabeled instance x_i is assigned a randomly selected label $\hat{y}_i \in \hat{\mathcal{Y}}$ to construct demonstrations $\mathcal{D}$:

$$\mathcal{D} = \{(x_1, \hat{y}_1), (x_2, \hat{y}_2), ..., (x_k, \hat{y}_k) | x_i \in N_k, \hat{y}_i \in \hat{\mathcal{Y}}\} \tag{2}$$

These synonyms are crafted to be semantically similar to 'real' and 'fake' but diverse enough to prevent direct replication of these terms by the LLM. By employing semantically rich alternatives, we ensure that the LLM engages more deeply with the content, activating its detection capabilities while preventing the copy effect.

Knowledge Retrieval Retrieval-Augmented methods have shown promising effects in many knowledge-intensive tasks as they allow for continuous knowledge updates and integration of domain-specific information (Lewis et al. 2020; Chang et al. 2024; Zhang, Zhang, and Pan 2022). To enable LLM to have a more detailed understanding of emergent news events and entities, we utilize wikipedia, which is continuously updated and widely used in knowledge-intensive tasks, as an external knowledge base to provide factual knowledge retrieval.

We utilize a LLM as an agent to retrieve key entities $\{k_1, k_2, ..., k_n\}$ from news content and then use the Wikipedia API[2] to retrieve information about these key entities $\mathcal{K} = \{(k_1, i_1), (k_2, i_2), ..., (k_n, i_n)\}$. The retrieved information with the news articles are provided to the LLM,

[1]https://api.bing.microsoft.com/v7.0/news/search

[2]https://www.wikipedia.org/

Algorithm 1: Pseudo-code for **MRCD**

Input : Emergent Events $\{X_e^t\} = \{x_{e,i}^t\}_{i=K+1}^N$, LLM $\mathcal{L}$, SLM $\mathcal{S}$ initialized with Past Events $\{X_e^s, Y_e^s\} = \{x_{e,i}^s, y_{e,i}^s\}_{i=1}^K$, round = 1

output : Labeled Emergent Events $\{X_e^t\} = \{x_{e,i}^t, y_{e,i}^t\}_{i=K+1}^N$

```
/* Multi-round Learning                    */
if round == 1 then
    Obtain Demonstrations D by BM25 and assign
      pseudo labels from (W, C);
    Knowledge-Retrieval Module to obtain K;
    Inference by L and S to obtain (ŷ1, ŷ2);
    ŷ1 = argmax_{ŷ1∈Y} P(ŷ1|D, K, x)
    ŷ2 = argmax_{ŷ2∈Y} P(ŷ2|x) ;
    Selection module to obtain D_clean and D_noisy;
    round = round + 1;
end
for round ≤ N do
    Retrieve D for L and fine-tune S by D_clean;
    D_noisy inference by L and S to obtain (ŷ1, ŷ2);
    Selection module to update D_clean and D_noisy;
    round = round + 1;
end
if D_noisy ≠ ∅ then
    D_noisy inference by S to obtain ŷ2;
    Update D_clean as final labels;
end
```

enabling it to have the most up-to-date and accurate external knowledge assistance for understanding and judgment.

Once the demonstrations $\mathcal{D}$ and related knowledge $\mathcal{K}$ are determined, MRCD concatenates them with the content of the news article x as input feed into the LLM. After in-context learning by demonstrations, LLM provides the predicted labels $\widehat{y_1}$ for x supplemented with related knowledge:

$$\widehat{y_1} = argmax_{\widehat{y_1} \in \mathcal{Y}} P(\widehat{y_1} | \mathcal{D}, \mathcal{K}, x) \tag{3}$$

The predicted labels $\widehat{y_1}$ from LLM and the news articles x are passed to the SLM for further data filtering and labeling.

Data Selection Module

As the news articles with pseudo labels $(x, \widehat{y_1})$ provided by the LLM are passed to the initialized SLM, the SLM directly infers pseudo labels $\widehat{y_2}$ for the news article x. In semi-supervised learning, using pseudo labels with high confidence as true labels for further training is a widely used approach (Rizve et al. 2020; Li et al. 2021). Inspired by this, we design a filtering mechanism that leverages the pseudo labels of both LLM and SLM to filter cleaner labels simultaneously. All unlabeled news $\{(x, \widehat{y_1}, \widehat{y_2}), x \in \mathcal{X}\}$ are divided into clean data pool D_{clean} and noisy data pool D_{noisy} by Equation 4, where $p(\widehat{y_2})$ denotes the output probability by SLM and ω denotes the confidence thresholds:

$$\begin{aligned} D_{clean} &= \{(x_i, y_i) \mid \widehat{y_1} = \widehat{y_2} \; and \; p(\widehat{y_2}) \geq \omega\} \\ D_{noisy} &= \{(x_i) \mid \widehat{y_1} \neq \widehat{y_2} \; or \; p(\widehat{y_2}) < \omega\} \end{aligned} \tag{4}$$

Category	Method	Pheme				Twitter16			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
SLM	RoBERTa	0.714	0.777	0.695	0.734	0.644	0.649	0.641	0.645
	EANN	0.744	0.738	0.745	0.741	0.641	0.621	0.741	0.676
	M^{3}FEND	0.746	0.747	0.746	0.746	0.642	0.608	**0.816**	0.697
	FTT	0.754	0.748	0.764	0.756	0.651	0.649	0.720	0.683
LLM (zero-shot)	Llama2-7B	0.505	0.498	0.697	0.581	0.496	0.501	0.494	0.498
	Llama3-8B	0.535	0.518	0.770	0.620	0.562	0.574	0.531	0.552
	GPT3.5	0.503	0.500	0.714	0.586	0.583	0.585	0.571	0.578
LLM (few-shot)	Llama2-7B	0.528	0.511	0.894	0.650	0.590	0.598	0.584	0.591
	Llama3-8B	0.549	0.524	**0.961**	0.679	0.622	0.607	0.717	0.658
	GPT3.5	0.520	0.507	0.850	0.635	0.621	0.609	0.705	0.653
LLM+SLM	ARG	0.743	0.741	0.779	0.760	0.705	0.698	0.710	0.704
MRCD	Llama2+RoBERTa	0.772	0.765	0.775	0.770	0.732	0.717	0.619	0.664
	GPT3.5+RoBERTa	0.781	0.735	0.821	0.778	0.768	0.752	0.734	0.743
	Llama3+RoBERTa	0.788	0.700	0.900	0.786	0.772	0.765	0.775	0.770
	Llama3+FTT	**0.814**	**0.788**	0.841	**0.814**	**0.794**	**0.768**	0.782	**0.774**
Improvements	*Impr. RoBERTa*	*+7.4%*	*/*	*+20.5%*	*+5.2%*	*+12.8%*	*+11.6%*	*+13.4%*	*+12.5%*
	Impr. FTT	*+6.0%*	*+4.0%*	*+7.7%*	*+5.8%*	*+14.3%*	*+11.9%*	*+13.2%*	*+12.1%*

Table 1: Performance of baselines and MRCD. Best results are in **bold** and second best results are underlined.

Multi-Round Learning

The initial classification of datasets into D_{noisy} and D_{clean} is only the beginning. Following this initial differentiation, a multi-round learning strategy is applied to further refine and enhance the quality of classifications, facilitating the dynamic adjustment of both the LLM and the SLM to capture emergent news efficiently.

Starting from the second round, these data in D_{noisy} is re-evaluated as unlabeled data and also undergoes the process of two-stage retrieval and data selection. Distinct from the initial round where demonstrations are primarily retrieved from external sources, in subsequent rounds, demonstrations $\mathcal{D}^{'}$ are drawn from the previously established D_{clean} utilizing the BM25 algorithm:

$$\begin{aligned} N_k^{'} &= \text{BM25}_{\text{top-k}}(x, D_{clean}) \\ \mathcal{D}^{'} &= \{(x_1, \hat{y}_1), (x_2, \hat{y}_2), ..., (x_k, \hat{y}_k) | x_i \in N_k^{'}\} \end{aligned} \tag{5}$$

In addition, the SLM will also use the data with clean labels for fine-tuning, further enhancing its ability to detect emergent news events. Apart from the demonstration retrieval module and SLM's fine-tuning, the rest modules of MRCD remain the same as in the first round. Through iterative rounds, we employ an iterative re-labeling strategy where each piece of data in D_{noisy} is reevaluated, and those meeting a specific confidence criterion are transferred to D_{clean}:

$$D_{clean}^{new} = \{(x_i, y_i) | x_i \in D_{noisy} \; and \; p(\widehat{y_2}) > \omega\} \cup D_{clean} \tag{6}$$

The newly validated instances in D_{clean}^{new} are then used as both training data for the SLM and demonstrations for the LLM in the upcoming iterations. When reaching the threshold $\mathcal{N}_{th}$ round, the remaining samples in D_{noisy} are directly predicted by the SLM $\mathcal{S}$ as the final judgment labels, thus completing the iterative learning cycle and solidifying the dataset classification. The overall pipeline is depicted in Algorithm 1, providing a detailed visualization of the multi-round learning.

Experiments

Datasets

To fairly evaluate the performance of the proposed model MRCD, we conduct experiments on datasets collected from real-world social media, namely Twitter16 (Detection and visualization of misleading content on Twitter 2018) and Pheme (Zubiaga, Liakata, and Procter 2017). Twitter16 and Pheme directly crawled news articles from trending events on Twitter, with each piece of data labeled with the associated event. Therefore, in Pheme, we use the latest occurring events germanwings-crash as the test set. For Twitter16, we directly use the events in the test set for evaluation.

Baselines

We select the following methods as baselines and divide them into three categories: SLM methods, LLM models and SLM+LLM methods. **For the SLM Methods. RoBERTa** (Liu et al. 2019) is an extension of BERT (Devlin et al. 2018) which employs dynamic masking strategies and larger batch sizes during pre-training for natural language understanding. **EANN** (Wang et al. 2018) firstly utilizes a discriminator to derive event-invariant features for multi-domain fake news detection. **M^{3}FEND** (Zhu et al. 2022) proposes a memory-guided multi-view framework to address the problem of domain shift and domain labeling incompleteness. **FTT** (Hu et al. 2023) adapts the model to future data by forecasting the temporal distribution patterns of news data. **For the LLMs. Llama2** (Touvron et al. 2023) is an auto-regressive, decoder-only large language model based on the Transformer architecture. **Llama3** uses the group query attention and a tokenizer with 128K words, based on Llama2.

Round	Model	Accuracy	Recall	Precision	F1-Score	Clean Pool	Noisy Pool
Round 0	LLM(zero-shot)	0.562	0.574	0.531	0.552	0	0
	LLM(few-shot)	0.622	0.607	0.717	0.658	0	0
Round 1	SLM	0.644	0.650	0.641	0.645	637	789
	LLM	0.604	0.588	0.722	0.650	637	789
Round 2	SLM	0.678	0.662	0.704	0.682	728	698
	LLM	0.674	0.639	0.704	0.678	728	698
Round 3	SLM	0.772	0.765	0.775	0.770	763	663
	LLM	0.665	0.661	0.739	0.689	763	663
Round 4	SLM	0.743	0.713	0.791	0.751	798	628
	LLM	0.663	0.658	0.692	0.674	798	628
Round 5	SLM	0.711	0.703	0.803	0.749	813	613
	LLM	0.667	0.653	0.695	0.671	813	613

Table 2: Analysis on multi-round learning.

GPT3.5 is a large language model, which has shown a surprising ability to solve NLP tasks without finetuning. We conduct zero-shot and few-shot experiments on these LLMs. In the few-shot experiments, we employed a 2-shot setup, wherein each LLM is provided with two real news samples and two fake news samples from test set as demonstrations.

LLM+SLM Methods. ARG (Hu et al. 2024) designs an adaptive rationale guidance network for fake news detection, in which SLMs selectively acquire information from LLM's rationales to enhance detection ability.

Implementation Details

Since we focus more on demonstrating the effectiveness of our framework MRCD, we use a simple pre-trained RoBERTa model (Liu et al. 2019) and an advanced FTT (Hu et al. 2023) model as the SLM. To verify the capabilities of different LLMs and their impact on MRCD, we select Llama2-7B, Llama3-8B and GPT-3.5 as the LLMs. To make a fair comparison, we replace the text feature extractors in the baseline SLMs architectures with a pre-trained RoBERTa. We set confidence threshold ω to 0.8, batch size to 32, round threshold $\mathcal{N}$ to 3, number of demonstrations k to 4. Our proposal is trained on 4 NVIDIA 3090 GPUs. The AdamW with a weight decay of 1e-4 is used as the optimizer and the initial learning rate is set to 1e-3.

Experimental Results

Table 1 illustrates the comparison between our approach MRCD and the baseline methods across two datasets. The three LLMs do not achieve satisfactory results in both zero-shot and few-shot settings, indicating that directly using LLMs for emergent fake news detection is not advisable. However, the application of all three LLMs contributes significantly to the success of MRCD. The Llama3+FTT model achieves the best results, demonstrating that the intrinsic capabilities of the LLM are crucial within our framework. MRCD(Llama3+RoBERTa) achieves an increase in accuracy of 7.4% and 12.8%, and an increase in f1-score of 5.2% and 12.5% compared to directly applying the SLM RoBERTa on Pheme and Twitter16 respectively. This demonstrates that MRCD significantly enhances the ability of SLMs to detect emergent news events by effectively utilizing the collaborative paradigm between LLMs and SLMs. Furthermore, despite using only a standard RoBERTa model as our SLM rather than a specialized fake news detection model, MRCD still outperforms existing SOTA SLMs for emergent fake news detection and ARG, which fully demonstrates the effectiveness of our approach. Notably, the chosen SLM RoBERTa in our framework can be replaced with other more advanced models specifically designed for emergent fake news detection. For instance, MRCD(Llama3+FTT) also proves the detection accuracy significantly compared to FTT, further demonstrating the universality of our framework as it can accommodate any more advanced SLM and LLM to improve the detection ability.

Multi-round Learning Analysis

To verify the effectiveness of multi-round learning, we conduct experiments on the inference results of both the LLM and SLM after each round of learning. Additionally, we record the changes in the number of samples in the clean data pool D_{clean} and noisy data pool D_{noisy} after each round. Here we use Llama3-8B as the LLM, RoBERTa as the SLM and conduct experiments on Twitter16 to investigate the changes. The results are shown in Table 2.

In the first round, after undergoing knowledge retrieval from external knowledge bases and demonstration retrieval from online engines and news corpus, the LLM shows significant improvement compared to its zero-shot inference results. This demonstrates that our two-stage retrieval effectively enhances the LLM's detection capabilities, achieving performance close to that of few-shot settings even without labeled samples. In the second and third round, benefiting from the clean samples used as demonstrations for the LLM and finetuning the SLM, both models show significant improvement in detection capabilities compared to the first round. Additionally, data from the noisy pool D_{noisy} gradually transitioned to the clean pool D_{clean}. This demonstrates that our multi-round learning approach effectively leverages pseudo-labeled samples to enhance the generalization capabilities of both LLM and SLM for emergent news events, thereby improving the detection accuracy and reliability of the pseudo labels. However, starting from the fourth round,

Variant Models	Pheme				Twitter16			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
MRCD(Llama3-Based)	0.788	0.700	0.900	0.786	0.772	0.765	0.775	0.770
w/o Demonstration	0.762	0.691	0.920	0.774	0.680	0.808	0.675	0.735
w/o Search Engine	0.774	0.712	0.850	0.771	0.763	0.754	0.761	0.757
w/o News Corpus	0.772	0. 691	0.883	0.778	0.758	0.765	0.754	0.759
w/o Knowledge	0.769	0.695	0.831	0.752	0.761	0.762	0.734	0.748
w/o Multi-Round	0.717	0.702	0.789	0.741	0.678	0.662	0.704	0.682

Table 3: Ablation Study on Pheme and Twitter16.

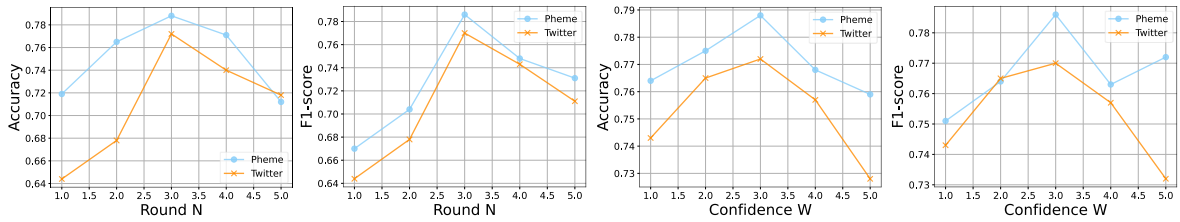


Figure 3: Hyper-parameter sensitivity analysis of $\mathcal{N}$ and ω.

the performance of the SLM experiences a noticeable decline. This is due to an increased number of noisy samples D_{noisy} contaminating the clean sample pool D_{clean}, which are then used for finetuning the SLM. In contrast, the LLM is less affected because in-context learning focuses more on the content of the demonstrations rather than the accuracy of the labels, which further supports the feasibility of extracting content from external sources and using pseudo-labels as demonstrations.

Ablation Study

To investigate the role of each module in MRCD, we conduct ablation experiments for five modules, as shown in Table 3. "w/o Demonstration" denotes the implementation without retrieving demonstrations from online search engines or news corpus for in-context learning. We further conduct experiments with "w/o Search Engine" and "w/o News Corpus" separately to further validate the roles of these two demonstration sources. "w/o Knowledge" represents the model results after removing the knowledge retrieval module. "w/o Multi-Round" denotes the model not utilizing multi-round learning but using the clean data D_{clean} from the first round to finetune the SLM and make the prediction.

When we remove the retrieval of demonstrations from the external sources, MRCD exhibits a significant decrease in performance across two datasets, especially on the Twitter dataset. This underscores the importance of providing demonstrations for the LLM through in-context learning. Furthermore, to further validate the impact of search engines and news corpus, we conduct ablation experiments on each source separately. The results indicate that eliminating either source led to d decline in model performance, highlighting the necessity of leveraging diverse external sources as demonstrations providers. To validate the impact of knowledge retrieval, we conduct an ablation study by removing the knowledge retrieval component. We observe that the model's performance declined across the two datasets, indicating that knowledge retrieval is crucial for enhancing the model's effectiveness. Finally, we also evaluate the performance without employing multi-round learning. We find that in the absence of multi-round learning, the model struggles to effectively utilize the pseudo-labeled data to adapt to emergent news events, resulting in suboptimal performance.

Parameter Sensitivity Analysis

Here we conduct hyper-parameter sensitivity analysis on the weights of two parameters: the threshold $\mathcal{N}$ of multi-round learning and the confidence thresholds ω of data selection. We observe that when the threshold $\mathcal{N}$ of multi-round learning is under 3, MRCD does not perform well in detection because it cannot extract enough clean samples in D_{clean} to fine-tune the SLM and provide demonstrations for the LLM's in-context learning. Conversely, when the threshold $\mathcal{N}$ is larger than 3, too many noisy samples may be incorporated into the clean pool D_{clean}, leading to a decline in the SLM's judgment capability. The same applies to confidence threshold ω: when ω is less than 0.8, the clean pool D_{clean} contains too much noisy samples, reducing the SLM's performance. On the other hand, when ω is larger than 0.8, there are too few samples to fine-tune the SLM and provide demonstrations for the LLM, resulting in poor accuracy as both models cannot adequately adapt to new news events.

Conclusion

Emergent fake news detection primarily faces the challenges of inconsistent distribution of emerging data and the lack of annotation. To address these challenges, in this paper we propose a multi-round collaboration framework between LLM and SLM for emergent fake news detection, dubbed MRCD. We propose a two-stage retrieval module, a data selection module, and a multi-round collaboration module to enhance detection capability in unsupervised emergent news events settings. Extensive experiments on two real-world datasets have proven the effectiveness of our model MRCD.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62272025 and No.U22B2021), and the Fund of the State Key Laboratory of Software Development Environment.

References

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Chen, C.; and Shu, K. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Chen, J.; Lu, Y.; Lin, H.; Lou, J.; Jia, W.; Dai, D.; Wu, H.; Cao, B.; Han, X.; and Sun, L. 2023. Learning In-context Learning for Named Entity Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13661–13675.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Cheng, D.; Huang, S.; Bi, J.; Zhan, Y.; Liu, J.; Wang, Y.; Sun, H.; Wei, F.; Deng, W.; and Zhang, Q. 2023. UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. In *EMNLP*, 12318–12337.

Detection; and visualization of misleading content on Twitter. 2018. Boididou, Christina and Papadopoulos, Symeon and Zampoglou, Markos and Apostolidis, Lazaros and Papadopoulou, Olga and Kompatsiaris, Yiannis. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gangireddy, S. C. R.; P, D.; Long, C.; and Chakraborty, T. 2020. Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM conference on hypertext and social media*, 75–83.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gruver, N.; Finzi, M.; Qiu, S.; and Wilson, A. G. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI*, volume 38, 22105–22113.

Hu, B.; Sheng, Q.; Cao, J.; Zhu, Y.; Wang, D.; Wang, Z.; and Jin, Z. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 116–125.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *ACL*, 754–763.

Khattab, O.; Santhanam, K.; Li, X. L.; Hall, D.; Liang, P.; Potts, C.; and Zaharia, M. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, C.; Pang, B.; Liu, Y.; Sun, H.; Liu, Z.; Xie, X.; Yang, T.; Cui, Y.; Zhang, L.; and Zhang, Q. 2021. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 223–232.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *DASFAA*, 163–179. Springer.

Li, C.; Wang, S.; Yu, P. S.; Zheng, L.; Zhang, X.; Li, Z.; and Liang, Y. 2018. Distribution distance minimization for unsupervised user identity linkage. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 447–456.

Li, X.; and Qiu, X. 2023. Finding Support Examples for In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6219–6235.

Lin, X. V.; Chen, X.; Chen, M.; Shi, W.; Lomeli, M.; James, R.; Rodriguez, P.; Kahn, J.; Szilvasy, G.; Lewis, M.; et al. 2023. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, W. B.; Carin, L.; and Chen, W. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 100–114.

Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024. FakeNewsGPT4: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs. *arXiv preprint arXiv:2403.01988*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyu, X.; Min, S.; Beltagy, I.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2304–2317.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Olan, F.; Jayawickrama, U.; Arakpogun, E. O.; Suklan, J.; and Liu, S. 2022. Fake news on social media: the impact on society. *Information Systems Frontiers*, 1–16.

Przybyla, P. 2020. Capturing the style of fake news. In *AAAI*, volume 34, 490–497.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations*.

Shu, K.; Mahudeswaran, D.; Wang, S.; and Liu, H. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *AAAI*, volume 14, 626–637.

Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *AAAI*, volume 35, 557–565.

Su, J.; Cardie, C.; and Nakov, P. 2023. Adapting fake news detection to the era of large language models. *arXiv preprint arXiv:2311.04917*.

Sun, M.; Zhang, X.; Ma, J.; Xie, S.; Liu, Y.; and Philip, S. Y. 2023a. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

Sun, X.; Dong, L.; Li, X.; Wan, Z.; Wang, S.; Zhang, T.; Li, J.; Cheng, F.; Lyu, L.; Wu, F.; et al. 2023b. Pushing the limits of chatgpt on nlp tasks. *arXiv preprint arXiv:2306.09719*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. *arXiv preprint arXiv:2402.10426*.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wu, J.; and Hooi, B. 2023. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2582–2593.

Wu, Z.; Wang, Y.; Ye, J.; and Kong, L. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In *ACL*, 1423–1436.

Ye, J.; Wu, Z.; Feng, J.; Yu, T.; and Kong, L. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, 39818–39833. PMLR.

Yin, S.; Zhu, P.; Wu, L.; Gao, C.; and Wang, Z. 2024. GAMC: An Unsupervised Method for Fake News Detection Using Graph Autoencoder with Masking. In *AAAI*, volume 38, 347–355.

Zhang, B.; Zhang, X.; Huang, F.; Lu, M.; and Ma, S. 2022. Deep Kernel Network Embedding. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5710–5723.

Zhang, J.; Li, Z.; Das, K.; and Kumar, S. 2024a. Interactive Multi-fidelity Learning for Cost-effective Adaptation of Language Model with Sparse Human Supervision. *Advances in Neural Information Processing Systems*, 36.

Zhang, L.; Zhang, X.; Li, C.; Zhou, Z.; Liu, J.; Huang, F.; and Zhang, X. 2024b. Mitigating social hazards: Early detection of fake news via diffusion-guided propagation path generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2842–2851.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11676–11684.

Zhang, L.; Zhang, X.; Zhou, Z.; Huang, F.; and Li, C. 2024c. Reinforced Adaptive Knowledge Learning for Multimodal Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16777–16785.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Wang, S.; Philip, S. Y.; and Li, C. 2024d. Early Detection of Multimodal Fake News via Reinforced Propagation Path Generation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J. B.; Zhang, Y.; Xie, X.; Wang, H.; and Kim, S. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 168–176.

Zhao, J.; Li, C.; Wen, Q.; Wang, Y.; Liu, Y.; Sun, H.; Xie, X.; and Ye, Y. 2021. Gophormer: Ego-graph transformer for node classification. *arXiv preprint arXiv:2110.13094*.

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

Zhou, Z.; Zhang, X.; Zhang, L.; Liu, J.; Zhang, X.; and Li, C. 2024. FineFake: A Knowledge-Enriched Dataset for Fine-Grained Multi-Domain Fake News Detecction. *arXiv preprint arXiv:2404.01336*.

Zhu, Y.; Sheng, Q.; Cao, J.; Nan, Q.; Shu, K.; Wu, M.; Wang, J.; and Zhuang, F. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Zubiaga, A.; Liakata, M.; and Procter, R. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, 109–123. Springer.