Robust Misinformation Detection by Visiting Potential Commonsense Conflict

Bing Wang[1,2], **Ximing Li**[1,2*], **Changchun Li**[1,2], **Bingrui Zhao**[1,2], **Bo Fu**[3]
Renchu Guan[1,2] and **Shengsheng Wang**[1,2*]
[1]College of Computer Science and Technology, Jilin University
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of the MoE, Jilin University
[3]School of Computer Science and Artificial Intelligence, Liaoning Normal University
{wangbing1416, liximing86, changchunli93}@gmail.com, fubo@lnnu.edu.cn, {guanrenchu, wss}@jlu.edu.cn

Abstract

The development of Internet technology has led to an increased prevalence of misinformation, causing severe negative effects across diverse domains. To mitigate this challenge, Misinformation Detection (MD), aiming to detect online misinformation automatically, emerges as a rapidly growing research topic in the community. In this paper, we propose a novel plug-and-play augmentation method for the MD task, namely Misinformation Detection with Potential Commonsense Conflict (MD-PCC). We take inspiration from the prior studies indicating that fake articles are more likely to involve commonsense conflict. Accordingly, we construct commonsense expressions for articles, serving to express potential commonsense conflicts inferred by the difference between extracted commonsense triplet and golden ones inferred by the well-established commonsense reasoning tool COMET. These expressions are then specified for each article as augmentation. Any specific MD methods can be then trained on those commonsense-augmented articles. Besides, we also collect a novel commonsense-oriented dataset named *CoMis*, whose all fake articles are caused by commonsense conflict. We integrate MD-PCC with various existing MD backbones and compare them across 4 public benchmark datasets and *CoMis*. Empirical results demonstrate that MD-PCC can consistently outperform the existing MD baselines.

1 Introduction

Over the past decades, many social media platforms *e.g.,* Twitter and Weibo, become the mainstream avenue to share information among human beings in daily life. Unfortunately, these platforms eventually afford convenience for the dissemination of various misinformation such as fake news and rumors [Vosoughi *et al.*, 2018; van der Linden, 2022]. To reduce the negative effect of misinformation, how to detect them effectively and efficiently becomes the primary task in this endeavor. Accordingly, the emergent topic of

[*]Corresponding authors

❶ **Article**: The body will produce toxins at any time, and if they accumulate too much, you will get sick. <u>Drinking more juice</u> will help to <u>eliminate toxins</u>.   **Veracity label**: *Fake*
❷ **Article**: <u>Meat floss</u> is made of <u>cotton</u>. This was discovered by my niece's mother-in-law. Moms, please pay attention.   **Veracity label**: *Fake*

Table 1: Real-world misinformation examples with commonsense conflict. The text fragments implying commonsense conflict are underlined. Human beings are more likely to identify these articles contain misinformation owing to the commonsense conflicts.

Misinformation **D**etection (**MD**) has recent drawn increasing attention from the natural language process community [Ma *et al.*, 2016; Zhang *et al.*, 2021; Sheng *et al.*, 2022; Hu *et al.*, 2023; Wang *et al.*, 2024a].

Generally, cutting-edge MD works employ a variety of deep learning techniques to learn the potential semantic correlation between online articles and their corresponding veracity labels, *e.g.,* real and fake [Ma *et al.*, 2016; Zhang *et al.*, 2021; Hu *et al.*, 2023; Zhang *et al.*, 2024]. For example, most MD arts concentrate on designing various models to incorporate external features, *e.g.,* entity-based embeddings of named entities in an article and their corresponding descriptions [Dun *et al.*, 2021; Hu *et al.*, 2021], domain information for adapting MD models across multiple domains [Nan *et al.*, 2022], and emotional signals to enhance MD models by learning potential emotional patterns [Zhang *et al.*, 2021].

Despite the success of learning the pattern between articles and veracity labels from data, we are particularly interested in, as complicated phenomenons, **how do human beings identify misinformation?** Recent psychological and sociological studies partially offer a certain kind of answer as human beings naturally distinguish misinformation by referring to their pre-existing commonsense knowledge [Lewandowsky *et al.*, 2012; Scheufele and Krause, 2019]. In certain scenarios, articles with misinformation are more likely to involve **commonsense conflict**, and human beings will identify misinformation by leveraging, at least referring to, such conflict involved, as examples illustrated in Table 1.

To identify misinformation by simulating the way of human thinking regarding commonsense conflict, the primary problem is how to measure and express them for given articles. Accordingly, we propose a novel plug-and-play augmentation method for the MD task, namely **M**isinformation **D**etection with **P**otential **C**ommonsense **C**onflict (**MD-PCC**). Specifically, we propose to measure the commonsense conflicts of articles by the difference between the extracted commonsense triplet and the golden triplet inferred by the well-established commonsense reasoning tool [Bosselut *et al.*, 2019; Hwang *et al.*, 2021], and use those triplets to specify a predefined *commonsense template* as *commonsense expressions* to express the potential commonsense conflicts. For each article, we integrate it with its corresponding specific commonsense expression to form an augmented one, named *commonsense-augmented article*. Given those augmented articles, one can build effective detectors by any existing MD methods and backbones.

For empirical evaluations, we employ 4 public benchmark datasets *GossipCop* [Shu *et al.*, 2020], *Weibo* [Sheng *et al.*, 2022], *PolitiFact* [Shu *et al.*, 2020] and *Snopes* [Popat *et al.*, 2017]. Additionally, we further collect a new **C**ommonsense-**o**riented **Mis**information benchmark datasets, named ***CoMis***, whose all fake articles are caused by commonsense conflict. We integrate MD-PCC with various existing MD backbones and compare them across public benchmark datasets and *CoMis*. Empirical results demonstrate that MD-PCC can consistently outperform the existing MD baselines. The source code and data of MD-PCC are released in the repository https://github.com/wangbing1416/MD-PCC.

The primary contributions of this paper can be summarized as the following three-folds:

- We propose a plug-and-play augmentation MD method, named MD-PCC, by expressing the potential commonsense conflict.
- We collect a new commonsense-oriented misinformation dataset, named *CoMis*, whose all fake articles are caused by commonsense conflict.
- We conduct experiments across both public benchmark datasets and *CoMis*, and empirical results indicate the effectiveness of MD-PCC.

2 Proposed MD-PCC Method

In this section, we briefly review the task definition of MD and prevalent commonsense reasoning methods. We then describe the proposed method MD-PCC in more detail.

2.1 Preliminaries

Task formulation of MD. Commonly, the basic goal of MD is to induce a detector $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ over a given training dataset $\mathcal{D}$, and use $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ to distinguish whether any unseen article is real or fake. We formally describe the dataset of N training samples as $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where each sample is composed of a raw article $\mathbf{x}_i$ and its corresponding veracity label $y_i \in \{0, 1\}$, *i.e.,* 0/1 indicating fake/real. With any specific detector $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$, it can be trained by optimizing the following objective with respect to $\boldsymbol{\theta}$:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \ell\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\right), \tag{1}$$

where $\ell(\cdot, \cdot)$ denotes the binary cross-entropy loss commonly.

Commonsense reasoning. Generally speaking, current commonsense reasoning methods aim to train a generative language model referring to the relation triplet $(\mathbf{s}, r, \mathbf{o})$, where $\mathbf{s}$ and $\mathbf{o}$ are the subject and object, respectively, and r is the relation between them. Given any subject-relation pair $(\mathbf{s}, r)$, a commonsense reasoning method can accurately predict the corresponding object $\mathbf{o}$. Typically, the methods are trained across the commonsense-oriented dataset ATOMIC$^{20}_{20}$ [Hwang *et al.*, 2021], which comprises a substantial collection of relation triplets. The typical commonsense-oriented relations of ATOMIC$^{20}_{20}$ include {`xNeed`, `xAttr`, `xReact`, `xEffect`, `xWant`, `xIntent`, `oEffect`, `oReact`, `oWant`, `isAfter`, `HasSubEvent`, `HinderedBy`}, representing the relations between specific events or human actions. Beyond these ones, these methods can also be generalized to a large knowledge base ConceptNet [Speer *et al.*, 2017], so as to capture the relations between entities including {`MadeOf`, `AtLocation`, `isA`, `Partof`, `HasA`, `UsedFor`}. To make notation simple, we use $\mathcal{R}$ to denote the set of all those relations captured by the commonsense reasoning methods.

2.2 Overview of MD-PCC

Basically, our MD-PCC is a plug-and-play augmentation method for the MD task. We take inspiration from the assumption that fake articles are more likely to involve commonsense conflict. Accordingly, we design a *commonsense template* to express the potential commonsense conflict measured by prevalent commensense reasoning methods and specify it for each original article as the augmentation. To be specific, the commonsense template is designed as

$$\mathbf{c} \oplus \mathbf{s} \oplus \Gamma(r) \oplus \hat{\mathbf{o}} \left[\oplus \text{“\textit{instead of}”} \oplus \mathbf{o} \right],$$

where $(\mathbf{s}, r, \mathbf{o})$ indicates the *representative commonsense triplet* extracted from the article, $\hat{\mathbf{o}}$ is the **golded** object corresponding to $(\mathbf{s}, r)$ generated by commonsense reasoning methods, and $\Gamma(r)$ denotes the original expression of r, *e.g.,* "*is made of*" is the original expression of `MadeOf`. We suppose that an article involves a commonsense conflict if $\mathbf{o} \neq \hat{\mathbf{o}}$, otherwise $\mathbf{o} = \hat{\mathbf{o}}$. Accordingly, we define that $\mathbf{c}$ will be specified by the adversative conjunction word "*However*" when $\mathbf{o} \neq \hat{\mathbf{o}}$; and by contrast, it will be specified by "*And*", and the text segment *"instead of"* $\oplus\, \mathbf{o}$ will be excluded.

With this commonsense template, for each article $\mathbf{x}_i$, we form its corresponding *commonsense expression* $\mathbf{e}_i$ by specifying $(\mathbf{s}_i, r_i, \mathbf{o}_i)$, $\hat{\mathbf{o}}_i$ and $\mathbf{c}_i$ with three stages: **commonsense triplet extraction**, **golden object generation**, and **commonsense expression construction**, respectively. Accordingly, we concatenate $\mathbf{x}_i$ and $\mathbf{e}_i$ as a commonsense-augmented article $\hat{\mathbf{x}}_i$. Given all commonsense-augmented samples $\{\hat{\mathbf{x}}_i, y_i\}_{i=1}^N$, we can formulate the following objective with respect to any specific detector $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \ell\left(\mathcal{F}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i), y_i\right), \; \hat{\mathbf{x}}_i = \mathbf{x}_i \oplus \mathbf{e}_i. \tag{2}$$

For clarity, the overall framework of MD-PCC is depicted in Fig. 1. In the following subsections, we will describe the three stages of generating commonsense expressions.

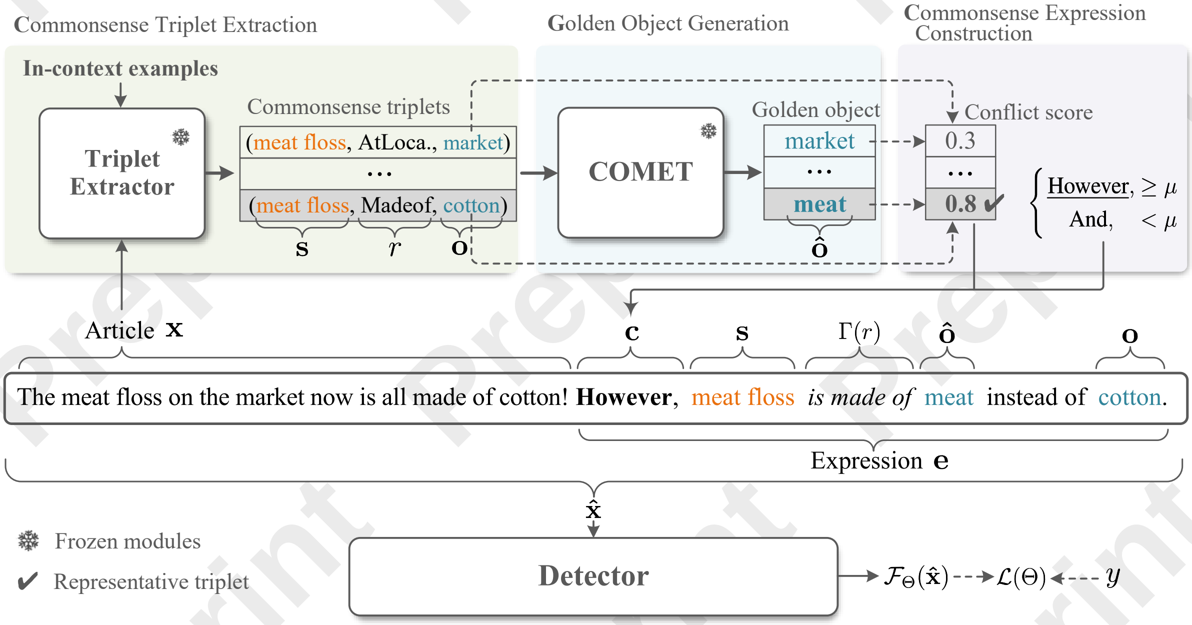

Figure 1: The overall framework of **MD-PCC**. Its basic idea is to construct a commonsense expression $\mathbf{e}$ and specify it as an augmentation. To achieve this, given an article $\mathbf{x}$, we input it and several in-context examples into a triplet extractor to extract commonsense triplets. Then, we generate corresponding golden objects for them using the commonsense tool. Finally, we calculate commonsense conflict scores for each pair of extracted and golden objects, and select one with the highest score, *e.g.,* 0.8, to construct the commonsense expression. In the framework, the parameters of the triplet extractor and COMET are frozen, and the detector will be optimized with Eq. (2).

2.3 Commonsense Triplet Extraction

In this stage, for each article $\mathbf{x}_i$, we extract a certain number of relevant commonsense triplets $\{(\mathbf{s}_i^\gamma, r_i^\gamma, \mathbf{o}_i^\gamma)\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$. To achieve this, we first screen all relations of $\mathcal{R}$ to extract all corresponding triplets $\{(\mathbf{s}_i^\gamma, r_i^\gamma, \mathbf{o}_i^\gamma)\}_{\gamma=1}^{|\mathcal{R}|}$ from $\mathbf{x}_i$ and then filter out the meaningless ones from them.

Specifically, we first extract $\{(\mathbf{s}_i^\gamma, r_i^\gamma, \mathbf{o}_i^\gamma)\}_{\gamma=1}^{|\mathcal{R}|}$ by prompting an existing LLM with the In-Context Learning (ICL) method [Brown *et al.*, 2020; Min *et al.*, 2022]. We design natural language queries $\{\mathcal{T}^\gamma\}_{\gamma=1}^{|\mathcal{R}|}$ for relations $\{r_i^\gamma\}_{\gamma=1}^{|\mathcal{R}|}$, *e.g.,* "*Extract entity1 and entity2 from the text where entity1 is made of entity2. Text:*" for the relation `MadeOf`. Accordingly, the formulation of in-context examples $\{\mathcal{I}_k^\gamma\}_{k=1}^{K}$ for the relation r^γ is delineated as:

$$\mathcal{I}_k^\gamma = \mathcal{T}^\gamma \oplus \mathbf{x}_k^\gamma \oplus \textit{entity1 is } \mathbf{s}_k^\gamma \textit{ and entity2 is } \mathbf{o}_k^\gamma,$$
$$\gamma \in \{1, 2, \cdots, |\mathcal{R}|\}, k \in \{1, 2, \cdots, K\}. \tag{3}$$

We collect K labeled examples for each relation to facilitate ICL. Then, we input both in-context examples and query $\mathcal{T}^\gamma \oplus \mathbf{x}_i$ into a triplet extractor $\mathcal{G}_\Phi(\cdot)$ specified by a pre-trained T5 model [Raffel *et al.*, 2020] to generate $\mathbf{s}_i^\gamma$ and $\mathbf{o}_i^\gamma$:

$$\mathbf{s}_i^\gamma, \mathbf{o}_i^\gamma \leftarrow \mathcal{G}_\Phi\left(\mathcal{I}_1^\gamma \oplus \cdots \oplus \mathcal{I}_K^\gamma \oplus \mathcal{T}^\gamma \oplus \mathbf{x}_i\right),$$
$$\gamma \in \{1, 2, \cdots, |\mathcal{R}|\}. \tag{4}$$

Because an article does not always contain all relations of $\mathcal{R}$, we filter out the meaningless ones. We design a filtering method based on the conditional generation logits. It follows the spirit that generative models always output lower probabilities for its generated uncertain word tokens, which can be evaluated by *perplexity* [Jurafsky, 2000; Lee *et al.*, 2021]. Specifically, we define the text generated by $\mathcal{G}_\Phi(\cdot)$ in Eq. (4) as $\mathbf{t}_i^\gamma = \{t_{i1}^\gamma, t_{i2}^\gamma, \cdots, t_{iL}^\gamma\}$, where L indicates the length of the text. And we remove the triplet $(\mathbf{s}_i^\gamma, r_i^\gamma, \mathbf{o}_i^\gamma)$, if

$$\sum_{j=1}^{L} \log P\left(t_{ij}^\gamma | t_{i<j}^\gamma; \Phi\right) > \epsilon,$$

where ϵ is a controllable hyper-parameter. After filtering, the set of commonsense relations for $\mathbf{x}_i$ is refined as $\overline{\mathcal{R}}_i \in \mathcal{R}$.

2.4 Golden Object Generation

Given $\{(\mathbf{s}_i^\gamma, r_i^\gamma)\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$, we generate their golden objects $\{\hat{\mathbf{o}}_i^\gamma\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$, which are aligned with real-world commonsense knowledge. To be specific, we feed each $(\mathbf{s}_i^\gamma, r_i^\gamma)$ into the prevalent commonsense tool $\mathcal{G}_\Pi(\cdot)$ [Bosselut *et al.*, 2019] to generate its golden object $\hat{\mathbf{o}}_i^\gamma$ as

$$\hat{\mathbf{o}}_i^\gamma \leftarrow \mathcal{G}_\Pi\left(\mathbf{s}_i^\gamma, r_i^\gamma\right), \ \gamma \in \{1, 2, \cdots, |\overline{\mathcal{R}}_i|\}. \tag{5}$$

We specially explain that because the prevalent commonsense reasoning tool has been pre-trained on a large-scale commonsense dataset ATOMIC$_{20}^{20}$, we treat $\hat{\mathbf{o}}_i^\gamma$ as the ground-truth knowledge of $(\mathbf{s}_i^\gamma, r_i^\gamma)$, *i.e.,* the corresponding golden object.

2.5 Commonsense Expression Construction

In this stage, we construct commonsense expression $\mathbf{e}_i$ by filling the commonsense template in Sec. 2.2 based on $\{(\mathbf{s}_i^\gamma, r_i^\gamma, \mathbf{o}_i^\gamma)\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$ and $\{\hat{\mathbf{o}}_i^\gamma\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$. Specifically, we first compute conflict scores $\{c_i^\gamma\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$ for each pair of $\mathbf{o}_i^\gamma$ and $\hat{\mathbf{o}}_i^\gamma$. We

Algorithm 1 Training summary of MD-PCC.

Input: Training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$; pre-trained language model $\mathcal{G}_{\Phi}(\cdot)$; commonsense reasoning tool $\mathcal{G}_{\Pi}(\cdot)$; commonsense relations $\mathcal{R}$; query templates $\{\mathcal{T}^{\gamma}\}_{\gamma=1}^{|\mathcal{R}|}$.
Output: detection model $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$; expressions $\{\mathbf{e}_i\}_{i=1}^{N}$.

```
for i = 1, 2, ..., N do
    C_i ← [ ],
    for r_i^γ in R do
        extract s_i^γ and o_i^γ with G_Φ(·) in Eq. (4),
        if Eq. (5) is not satisfied then
            ô_i^γ ← G_Π(s_i^γ, r_i^γ),
            calculate c_i^γ with Eq. (6), C_i ← c_i^γ.
        end if
    end for
    select {s_i, r_i, o_i} and ô_i with max(C_i),
    construct e_i with Eq. (7).
end for
train F_θ(·) with L in Eq. (2).
```

take inspiration from BARTSCORE [Yuan *et al.*, 2021], and present a new evaluation metric to compute the commonsense conflict score c_i^{γ} during the process that we input $\mathbf{s}_i^{\gamma}$ and r_i^{γ} into the commonsense reasoning tool $\mathcal{G}_{\Pi}(\cdot)$ with Eq. (5). The specific metric is as follows:

$$c_i^{\gamma} = -\sum\nolimits_{j=1}^{\overline{L}} \mathbf{o}_{ij}^{\gamma} \log \mathcal{P}\left(\hat{\mathbf{o}}_{ij}^{\gamma} \middle| \hat{\mathbf{o}}_{i<j}^{\gamma}; \Pi\right), \quad \gamma \in \{1, 2, \cdots, |\overline{\mathcal{R}}_i|\}, \tag{6}$$

where $\overline{L}$ denotes the length of the generated $\hat{\mathbf{o}}_{ij}^{\gamma}$.

Then, we select the highest conflict score c_i from the set of $\{c_i^{\gamma}\}_{\gamma=1}^{|\overline{\mathcal{R}}_i|}$, and denote its corresponding representative commonsense triplet and golden object as $\{\mathbf{s}_i, r_i, \mathbf{o}_i\}$ and $\hat{\mathbf{o}}_i$, respectively. Accordingly, we fill them into the commonsense template to obtain the expression as follows:

$$\mathbf{e}_i = \begin{cases} \text{“However”} \oplus \mathbf{s}_i \oplus \Gamma(r_i) \oplus \hat{\mathbf{o}}_i \oplus \\ \qquad \text{“instead of”} \oplus \mathbf{o}_i, \; c_i \geq \mu, \\ \text{“And”} \oplus \mathbf{s}_i \oplus \Gamma(r_i) \oplus \hat{\mathbf{o}}_i, \quad c_i < \mu, \end{cases} \tag{7}$$

where $\Gamma(\cdot)$ is the original expression for each relation, *e.g.,* "*is made of*" for the relation `MadeOf`. When $c_i \geq \mu$, we argue that the article $\mathbf{x}_i$ exists the commonsense conflict; otherwise, there is not. In summary, the training summary of MD-PCC is presented in Alg. 1.

3 Datasets

To evaluate the performance of MD-PCC, we conduct experiments by employing four public MD datasets *GossipCop* [Shu *et al.*, 2020], *Weibo* [Sheng *et al.*, 2022], *PolitiFact* [Shu *et al.*, 2020] and *Snopes* [Popat *et al.*, 2017]. Additionally, we also collect a new Chinese MD dataset, referred to as ***CoMis***, wherein all fake articles can be verified by leveraging commonsense conflict. We describe their details in the following section. For clarity, their statistics are shown in Table 2.

3.1 Prevalent MD Datasets

We evaluate the method with the following four MD datasets:

Dataset	# Train		# Val.		# Test	
	Fake	Real	Fake	Real	Fake	Real
Weibo	2,561	7,660	499	1,918	754	2,957
GossipCop	2,024	5,039	604	1,774	601	1,758
PolitiFact	1,224	1,344	170	186	307	337
Snopes	2,288	838	317	116	572	210
CoMis	560	440	170	125	162	123

Table 2: Statistics of prevalent FND datasets and *CoMis*.

- ***GossipCop*** and ***PolitiFact*** are English MD datasets sourced from *FakeNewsNet* [Shu *et al.*, 2020]. We divide *GossipCop* based on [Zhu *et al.*, 2022], which includes articles posted between 2000 and 2017 for training, with the test set consisting of articles from 2018. For *PolitiFact*, we adhere to its original dataset division.
- ***Weibo*** [Sheng *et al.*, 2022] is sourced from a Chinese social media platform, and we split articles published from 2010 to 2017 allocated for training and those from 2018 used for testing.
- ***Snopes*** [Popat *et al.*, 2017] is gathered from a well-known fact-checking website *snopes.com*. We split the dataset according to its original paper.

3.2 Our Collected *CoMis*

We collect a new commonsense-oriented MD dataset *CoMis* with the effort of human annotators. Table 2 provides the statistics of our newly constructed dataset *CoMis*. The dataset contains a total of 1,580 pieces of data entries, covering diverse domains. The domain most extensively represented in the dataset pertains to food safety.

Data source. Our MD data is sourced from two distinct channels: pre-existing MD datasets and external websites.

First, we select suitable data items from pre-existing datasets dedicated to fake news and rumor detection, *e.g., Weibo-16*, *Weibo-20*, and *Weibo-COVID19*. Specifically, *Weibo-16* [Ma *et al.*, 2016] comprises posts spanning from December 2010 to April 2014, and many duplications are meticulously filtered by [Zhang *et al.*, 2021]; *Weibo-20* [Zhang *et al.*, 2021] extends the temporal scope of *Weibo-16*, encompassing data from April 2014 to November 2018, and its labels are verified through NewsVerify[1]; *Weibo-COVID19* [Lin *et al.*, 2022] is collected during the surge of the COVID-19 pandemic, so all articles within this dataset are exclusively centered on COVID-19 topics.

To ensure completeness and timeliness, we also manually collect commonsense-oriented samples from two external websites. First, *Food Rumor*[2] is a Chinese rumor-refuting platform, which serves as a repository for misinformation and its corresponding verification, with a predominant focus on topics related to food safety, health science, and similar domains. Then, *Science Facts*[3] is another Chinese platform that specializes in disseminating science popularization content, covering subjects, *e.g.,* food safety and biological science.

[1]https://www.newsverify.com/
[2]http://www.xinhuanet.com/food/sppy/
[3]https://piyao.kepuchina.cn/

Method	Macro F1	Accuracy	Precision	Recall	F1_{real}	F1_{fake}	Avg.Δ
Dataset: *Weibo*							
EANN [Wang *et al.*, 2018]	76.53±0.52	84.62±0.30	76.75±0.63	76.07±1.14	90.43±0.25	62.41±1.12	-
+ MD-PCC (ours)	77.30±0.99*	85.88±0.50*	78.58±0.89*	76.29±0.89	91.25±0.32*	63.36±0.78*	**+0.98**
BERT [Devlin *et al.*, 2019]	75.64±0.41	84.13±0.67	75.58±1.09	75.79±0.74	90.02±0.52	61.26±0.59	-
+ MD-PCC (ours)	76.80±0.86*	84.62±0.92	76.32±1.41*	77.44±0.80*	90.26±0.67	63.35±1.16*	**+1.06**
BERT-EMO [Zhang *et al.*, 2021]	76.17±0.48	84.60±0.40	76.27±0.64	76.11±0.85	90.34±0.31	61.99±0.89	-
+ MD-PCC (ours)	77.03±1.21*	85.29±1.19*	77.50±1.00*	76.72±0.94*	91.53±0.80*	63.28±0.69*	**+0.98**
CED [Wu *et al.*, 2023]	76.42±1.55	85.51±1.32	77.92±0.87	75.70±0.63	90.72±0.91	62.42±1.40	-
+ MD-PCC (ours)	78.33±0.20*	86.59±0.51*	79.98±1.22*	77.13±1.11*	91.70±0.42*	64.96±0.63*	**+1.67**
DM-INTER [Wang *et al.*, 2024a]	76.29±0.42	84.59±0.33	76.23±0.51	76.39±0.87	90.31±0.27	62.26±0.84	-
+ MD-PCC (ours)	77.59±0.23*	85.80±0.72*	78.43±0.77*	77.32±0.74*	91.15±0.58*	64.13±0.64*	**+1.39**
Dataset: *GossipCop*							
EANN [Wang *et al.*, 2018]	78.59±0.84	84.47±0.66	80.37±1.46	77.42±1.36	89.80±0.55	67.39±1.59	-
+ MD-PCC (ours)	79.80±0.47*	85.08±0.35*	80.82±0.86	79.02±1.05*	90.12±0.32	69.48±0.99*	**+1.05**
BERT [Devlin *et al.*, 2019]	78.23±0.45	83.78±0.80	79.00±1.45	77.49±0.57	89.21±0.69	67.24±0.45	-
+ MD-PCC (ours)	79.10±0.46*	84.61±0.56*	80.32±1.10*	78.24±0.47*	89.85±0.45*	68.37±0.60*	**+0.92**
BERT-EMO [Zhang *et al.*, 2021]	78.42±0.47	83.92±0.39	79.15±0.73	77.10±1.01	89.67±0.59	67.23±1.03	-
+ MD-PCC (ours)	79.32±0.27*	84.68±0.66*	80.28±1.38*	78.63±0.67*	90.03±0.36	68.81±0.31*	**+1.04**
CED [Wu *et al.*, 2023]	78.33±0.40	83.77±0.68	78.85±1.26	77.94±0.25	89.17±0.57	67.49±0.25	-
+ MD-PCC (ours)	79.79±0.52*	85.52±0.31*	82.04±0.67*	78.23±0.84	90.54±0.22*	69.04±0.96*	**+1.60**
DM-INTER [Wang *et al.*, 2024a]	78.29±0.56	84.04±0.40	79.43±0.87	77.43±1.00	89.45±0.34	67.21±1.09	-
+ MD-PCC (ours)	79.76±0.42*	85.08±0.30*	80.85±0.75*	78.93±0.93*	90.13±0.28*	69.40±0.87*	**+1.38**
Dataset: *PolitiFact*							
BERT [Devlin *et al.*, 2019]	60.36±0.99	60.49±2.04	60.53±2.18	60.45±2.08	62.86±1.74	56.62±2.25	-
+ MD-PCC (ours)	61.92±0.68*	62.45±0.47*	62.46±0.39*	62.05±0.57*	66.29±0.46*	57.55±1.70*	**+1.90**
CED [Wu *et al.*, 2023]	61.75±0.54	61.86±0.50	61.79±0.51	61.77±0.54	63.56±0.90	59.94±1.23	-
+ MD-PCC (ours)	63.60±0.21*	63.87±0.34*	63.84±0.37*	63.63±0.23*	66.59±1.28*	60.61±1.05*	**+1.91**
DM-INTER [Wang *et al.*, 2024a]	60.85±1.96	61.23±1.77	61.23±1.71	60.97±1.81	64.15±1.56	57.54±1.57	-
+ MD-PCC (ours)	63.13±1.58*	63.37±1.51*	63.29±1.51*	63.14±1.55*	66.08±1.28*	60.17±1.17*	**+2.20**
Dataset: *Snopes*							
BERT [Devlin *et al.*, 2019]	62.74±0.78	72.15±1.74	64.36±2.03	62.14±0.70	43.56±1.71	81.91±1.58	-
+ MD-PCC (ours)	64.69±1.36*	73.42±1.89*	65.99±1.50*	64.14±1.20*	47.19±1.48*	82.19±1.43*	**+1.79**
CED [Wu *et al.*, 2023]	63.60±1.15	72.39±0.93	64.34±0.79	63.29±1.51	45.74±1.93	81.44±1.06	-
+ MD-PCC (ours)	66.41±1.32*	74.82±0.77*	67.46±1.02*	65.79±1.49*	49.61±1.58*	83.21±0.63*	**+2.75**
DM-INTER [Wang *et al.*, 2024a]	63.24±1.37	72.83±0.84	64.41±1.28	62.62±1.35	44.47±1.41	82.01±0.58	-
+ MD-PCC (ours)	65.79±1.34*	74.01±1.11*	66.51±1.39*	65.38±0.72*	49.06±1.86*	82.53±0.92	**+2.28**

Table 3: Experimental results of our MD-PCC on four prevalent datasets *Weibo*, *GossipCop*, *PolitiFact* and *Snopes*. The results marked by $*$ indicate that they are statistically significant than the baseline methods (p-value <0.05).

Annotation and post-process. The annotators are instructed to select and post-process the data items that can be verified using commonsense from the aforementioned data sources. For the data from pre-existing MD datasets, we preserve their veracity labels while systematically filtering any special symbols and website links from their content. For the data sourced from external websites, we collect fake claims in the rumor-refuting channels of these websites, and real claims from their science popularization channels. Meanwhile, we maintain a consistent average claim length of approximately 50, aligning with the standards set by existing datasets.

4 Experimental Results

In this section, we aim to empirically evaluate our proposed method MD-PCC, and answer the following questions:

- **Q1**: Can the proposed MD-PCC consistently improve the performance of misinformation detectors?
- **Q2**: Is MD-PCC sensitive to its hyper-parameters and primary components?
- **Q3**: Can the generated expression $\mathbf{e}$ expresses the commonsense conflict of the article?

4.1 Experimental Settings

Baselines. We evaluate our plug-and-play method MD-PCC across five prevalent MD approaches, including **EANN** [Wang *et al.*, 2018], **BERT** [Devlin *et al.*, 2019], **BERT-EMO** [Zhang *et al.*, 2021], the SOTA MD model **CED** [Wu *et al.*, 2023], and **DM-INTER** [Wang *et al.*, 2024a].

Implementation Details. In our experiments, we employ pre-trained language models $\text{FlanT5}_{\text{Large}}$[4] [Chung *et al.*, 2024] and $\text{mT5}_{\text{Large}}$[5] [Xue *et al.*, 2021] to extract common-

[4]https://huggingface.co/google/flan-t5-large.
[5]https://huggingface.co/google/mt5-large.

Method	Macro F1	Accuracy	Precision	Recall	$F1_{real}$	$F1_{fake}$	Avg.Δ
BERT [Devlin *et al.*, 2019]	88.70±0.53	89.02±0.56	88.90±0.69	88.54±0.42	88.22±0.67	90.60±0.55	-
+ MD-PCC (ours)	91.55±0.36*	91.71±0.35*	91.42±0.33*	91.78±0.48*	90.37±0.47*	92.72±0.34*	**+2.60**
CED [Wu *et al.*, 2023]	89.22±1.09	89.58±1.00	89.82±0.88	88.88±1.27	87.31±1.45	91.12±0.77	-
+ MD-PCC (ours)	91.69±1.11*	91.86±1.12*	91.61±1.25*	91.87±0.93*	90.51±1.16*	92.78±1.07*	**+2.40**
DM-INTER [Wang *et al.*, 2024a]	89.34±0.74	89.57±0.71	89.24±0.68	89.47±0.84	87.78±0.93	90.90±0.56	-
+ MD-PCC (ours)	91.61±0.96*	91.81±0.97*	91.63±0.92*	91.62±0.80*	90.31±1.00*	92.90±0.92*	**+2.26**

Table 4: Experimental results of our MD-PCC on our constructed datasets *CoMis*. The results marked by $*$ indicate that they are statistically significant than the baseline methods (p-value <0.05).

Method	F1	Acc.	Pre.	Rec.	$F1_{real}$	$F1_{fake}$
Dataset: *Weibo*						
CED	76.42	85.51	77.92	75.70	90.72	62.42
+ MD-PCC	78.33	86.59	79.98	77.13	91.70	64.96
w/o ICL	76.43	84.84	76.87	76.39	90.49	62.38
w/o **c**	77.22	85.33	77.55	76.65	90.76	63.43
w/o **o**	77.45	85.56	77.87	77.69	91.00	64.14
Dataset: *GossipCop*						
CED	78.33	83.77	78.85	77.94	89.17	67.49
+ MD-PCC	79.79	85.52	82.04	78.23	90.54	69.04
w/o ICL	78.40	83.93	79.15	77.80	89.33	67.46
w/o **c**	78.90	84.85	81.01	77.41	90.10	67.69
w/o **o**	79.27	84.65	80.13	78.54	89.83	68.71

Table 5: Ablative study of MD-PCC on two datasets *Weibo* and *GossipCop*. w/o represents without, and **c** and **o** are conjunctions and extracted objects in commonsense expressions, respectively.

sense triplets for the English and Chinese MD datasets, respectively. To generate golden objects, we use COMET-ATOMIC$^{20}_{20}$[6] [Hwang *et al.*, 2021] for English datasets and *comet-atomic-zh*[7] for Chinese datasets *Weibo* and *CoMis*.

During the training stage, we use an Adam optimizer with a learning rate of 7×10^{-5} for the BERT model in baseline methods. For the other modules such as the linear classifier, we use a learning rate of 1×10^{-4}, and the batch size is consistently fixed to 64. We also fix some other manual parameters empirically, such as K, ϵ, and μ to 5, 0.8, and 0.6, respectively. To avoid overfitting of detectors, we adopt an early stop strategy. This means that the training stage will stop when no better Macro F1 value appears for 10 epochs.

4.2 Main Results (Q1)

To answer Q1, Tables 3 and 4 report the performance outcomes of our method MD-PCC on two benchmark datasets and our constructed dataset, respectively. To mitigate the influence of randomness, we repeat each experiment five times using five different seeds $\{1, 2, 3, 4, 5\}$. The standard deviations of the five replicates are also illustrated in Tables 3 and 4. Overall, our MD-PCC method, which functions as a plug-in approach, can significantly and consistently improve the performance of the baseline models across all evaluation metrics. For example, on the *Weibo* dataset, our MD-PCC improves the overall F1 and fake news F1 scores by 1.91 and 2.54, respectively, compared to the current state-of-the-art MD method CED. Additionally, on *GossipCop*, it achieves improvements of 1.46 and 3.19 in macro F1 and precision scores. When we compare different MD datasets, we observe that MD-PCC performs better on *CoMis* than on the other Chinese dataset *Weibo* across most evaluation metrics. Specifically, when compared to the BERT baseline, MD-PCC improves its macro F1 and precision scores by 1.16 and 0.74 on *Weibo*, while it shows more significant improvements of 2.85 and 2.52 on *CoMis*. These results highlight the effectiveness of MD-PCC in incorporating commonsense knowledge to enhance the detection of knowledge-rich misinformation.

4.3 Ablative Study (Q2)

To investigate Q2, we implement ablative experiments to assess the effectiveness of key components in MD-PCC. Specifically, we conduct experiments on *Weibo* and *GossipCop*, and present three ablative versions of CED + MD-PCC as follows:

- **MD-PCC w/o ICL**: the version without ICL ($K = 0$) in the commonsense triplet extraction stage;
- **MD-PCC w/o c**: the version without conjunction words **c**, *e.g.,* "*However*", in commonsense expressions;
- **MD-PCC w/o o**: the version without "*instead of*" $\oplus$ **o** in commonsense expressions.

The ablative results are presented in Table 5. Generally, each ablative version exhibits a decreasing trend compared to MD-PCC, illustrating the contribution of each component in our model. The overall performance ranking of these ablative versions is w/o **o** >w/o **c** >w/o ICL. This ordering indicates that: (1) the direct impact of commonsense triplet extraction on the model's performance is significant, and in-context learning consistently enhances the extraction; (2) conjunction words **c** is more important than **o** in commonsense expressions. This is because misinformation detectors can effectively learn the pattern between conjunctions and veracity labels, *e.g.,* the pattern between "*However*" and *Fake*.

4.4 Case Study (Q3)

The goal of MD-PCC is to construct commonsense expressions that express the potential commonsense conflict. Therefore, we provide some representative cases in Table **??** to evaluate the generated expressions. Specifically, we select two cases from *CoMis* and translate them into English versions. We observe that (1) MD-PCC extracts commonsense triples accurately, even from relatively complex articles, *e.g.,* the second case; (2) MD-PCC can assign a higher conflict

[6]https://github.com/allenai/comet-atomic-2020.
[7]https://huggingface.co/svjack/comet-atomic-zh.

Article: Meat floss is made of cotton. This was discovered by my niece's mother-in-law. Moms, please pay attention.
Expression: However, meat floss is made of meatloaf instead of cotton.

	relation r	subject $\mathbf{s}$	object $\mathbf{o}$	gold object $\widehat{\mathbf{o}}$	conflict score c
❶	`MadeOf`	meat floss	cotton	meatloaf	**0.853**
❷	`IsA` / `HasA`	meat floss	cotton	crew meat / eat meat	0.728 / 0.835
❸	`AtLocation`	meat floss and cotton	-	-	-

Article: Everyone has been recommended "anti-blue light glasses" when they go shopping for glasses. Whether they are buying for themselves, these glasses seem to have become a must-have.Wearing it is good for your eyes and can even prevent myopia.
Expression: However, anti-blue light glasses show the effect on getting rid of blue light instead of preventing myopia.

	relation r	subject $\mathbf{s}$	object $\mathbf{o}$	gold object $\widehat{\mathbf{o}}$	conflict score c
❶	`isA`	anti-blue light glasses	glasses	protective eyeglasses	0.313
❷	`xEffect`	anti-blue light glasses	prevent myopia	get rid of blue light	**0.665**
❸	`HinderedBy`	PersonX has anti-blue light glasses	-	-	-

Table 6: Case study of MD-PCC on the dataset *CoMis*.

score to the triplet that does exist the commonsense conflict; (3) our presented filtering method in Sec. 2.3 can indeed filter out commonsense relations that do not exist in the article, *e.g.,* `AtLocation` in the first case.

5 Related Works

In this section, we briefly review the related literature about misinformation detection and commonsense reasoning.

5.1 Misinformation Detection

Misinformation, *e.g.,* fake news and rumors, has had a detrimental impact on society [Vosoughi *et al.*, 2018; Zhang *et al.*, 2023]. As a result, it has become increasingly important to identify and detect misinformation, which is referred to as misinformation detection. Specifically, most cutting-edge MD techniques focus on detecting misinformation based on its textual and multimodal content [Ying *et al.*, 2023; Wang *et al.*, 2024b], using advanced deep learning models [Ma *et al.*, 2016; Shu *et al.*, 2020; Wang *et al.*, 2024c; Xiao *et al.*, 2024]. These models often incorporate external features like knowledge bases [Dun *et al.*, 2021], emotional signals [Zhang *et al.*, 2021; Jiang *et al.*, 2024], and user feedback [Ma *et al.*, 2016; Lin *et al.*, 2023]. Meanwhile, some recent works have also explored strategies to leverage pre-trained large models for MD [Hu *et al.*, 2024; Chen and Shu, 2024; Nan *et al.*, 2024; Wan *et al.*, 2024].

In this study, we integrate commonsense knowledge into MD models. Prior to our work, certain research efforts have aimed to leverage knowledge graphs for the enhancement of MD models. These endeavors have primarily involved learning knowledge embeddings [Dun *et al.*, 2021; Sun *et al.*, 2022] or retrieving entity descriptions [Hu *et al.*, 2021; Jiang *et al.*, 2022]. In contrast to these approaches, our incorporation of commonsense knowledge aligns more closely with human reasoning and reactions. Meanwhile, we employ generative models for data augmentation explicitly, which obviates the need for extensive retrieval from large knowledge bases and reduces computational complexity.

5.2 Commonsense Bases and Reasoning

Commonsense knowledge bases, such as ConceptNet [Speer *et al.*, 2017], ATOMIC$^{20}_{20}$ [Hwang *et al.*, 2021], offer a valuable resource for direct reasoning with commonsense knowledge and have found applications in various academic topics, *e.g.,* machine translation [Liu *et al.*, 2023], question answering [Wang *et al.*, 2023; Chen *et al.*, 2023] and sarcasm detection [Min *et al.*, 2023]. Recently, especially within the context of Large Language Models (LLMs), the utilization of commonsense reasoning with LLMs has garnered significant attention [Liu *et al.*, 2023; Shen, 2024; Wang *et al.*, 2024d]. These works frequently treat commonsense knowledge as supplementary information or assess the presence of commonsense knowledge within LLMs.

6 Conclusion

In this paper, we aim to enhance MD models by uncovering commonsense conflicts. To achieve this goal, we propose a novel MD method named MD-PCC, designed to generate commonsense expressions for each article, explicitly expressing commonsense conflict existing inherent in articles, and leverage it to augment original articles. Specifically, the expression is constructed through a commonsense triplet extracted from the original article, the corresponding golden object, and a conjunction word. To obtain these components, we first prompt the pre-trained language model with in-context examples to extract triplets and filter out irrelevant triplets. Then, the commonsense tool is employed to generate their corresponding golden objects. Finally, a new metric is designed to measure the commonsense conflict, and the conjunction word is determined using this metric. Additionally, we also collect a new commonsense-oriented MD dataset, and extensive experimental results on the datasets are conducted and prove the effectiveness of our proposed MD-PCC.

Acknowledgements

We acknowledge support for this project from the National Key R&D Program of China (No.2021ZD0112501, No.2021ZD0112502), the National Natural Science Foundation of China (No.62276113), China Postdoctoral Science Foundation (No.2022M721321).

References

[Bosselut *et al.*, 2019] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *COLING*, pages 4762–4779, 2019.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Chen and Shu, 2024] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? In *ICLR*, 2024.

[Chen *et al.*, 2023] Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si, and Yin Zhang. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. In *Findings of ACL*, pages 13207–13224, 2023.

[Chung *et al.*, 2024] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 25:70:1–70:53, 2024.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[Dun *et al.*, 2021] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. KAN: knowledge-aware attention network for fake news detection. In *AAAI*, pages 81–89, 2021.

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

[Hu *et al.*, 2023] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *ACL: Industry Track*, pages 116–125, 2023.

[Hu *et al.*, 2024] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI*, pages 22105–22113, 2024.

[Hwang *et al.*, 2021] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, pages 6384–6392, 2021.

[Jiang *et al.*, 2022] Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029, 2022.

[Jiang *et al.*, 2024] Siqi Jiang, Zeqi Guo, and Jihong Ouyang. What makes sentiment signals work? sentiment and stance multi-task learning for fake news detection. *Knowledge-Based Systems*, 303:112395, 2024.

[Jurafsky, 2000] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[Lee *et al.*, 2021] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Towards few-shot fact-checking via perplexity. In *NAACL*, pages 1971–1981, 2021.

[Lewandowsky *et al.*, 2012] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

[Lin *et al.*, 2022] Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Guang Chen. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of NAACL*, pages 2543–2556, 2022.

[Lin *et al.*, 2023] Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. Zero-shot rumor detection with propagation structure via prompt learning. In *AAAI*, pages 5213–5221, 2023.

[Liu *et al.*, 2023] Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *ACL*, pages 15536–15550, 2023.

[Ma *et al.*, 2016] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, pages 3818–3824, 2016.

[Min *et al.*, 2022] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, pages 11048–11064, 2022.

[Min *et al.*, 2023] Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. Just like a human would, direct access to sarcasm augmented with potential result and reaction. In *ACL*, pages 10172–10183, 2023.

[Nan *et al.*, 2022] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. Improving fake news detection of influential domain via domain- and instance-level transfer. In *COLING*, pages 2834–2848, 2022.

[Nan *et al.*, 2024] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *CIKM*, pages 1732–1742, 2024.

[Popat *et al.*, 2017] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *WWW Companion*, pages 1003–1012, 2017.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:140:1–140:67, 2020.

[Scheufele and Krause, 2019] Dietram A Scheufele and Nicole M Krause. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669, 2019.

[Shen, 2024] Ke Shen. The generalization and robustness of transformer-based language models on commonsense reasoning. In *AAAI*, pages 23419–23420, 2024.

[Sheng *et al.*, 2022] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. Zoom out and observe: News environment perception for fake news detection. In *ACL*, pages 4543–4556, 2022.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.

[Sun *et al.*, 2022] Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. DDGCN: dual dynamic graph convolutional networks for rumor detection on social media. In *AAAI*, pages 4611–4619, 2022.

[van der Linden, 2022] Sander van der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28:460–467, 2022.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Wan *et al.*, 2024] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. DELL: generating reactions and explanations for llm-based misinformation detection. In *Findings of ACL*, pages 2637–2667, 2024.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: event adversarial neural networks for multi-modal fake news detection. In *KDD*, pages 849–857, 2018.

[Wang *et al.*, 2023] Wenya Wang, Vivek Srikumar, Hannaneh Hajishirzi, and Noah A. Smith. Elaboration-generating commonsense question answering at scale. In *ACL*, pages 1619–1635, 2023.

[Wang *et al.*, 2024a] Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. Why misinformation is created? detecting them by integrating intent features. In *CIKM*, pages 2304–2314, 2024.

[Wang *et al.*, 2024b] Bing Wang, Ximing Li, Changchun Li, Shengsheng Wang, and Wanfu Gao. Escaping the neutralization effect of modality features fusion in multimodal fake news detection. *Information Fusion*, 111:102500, 2024.

[Wang *et al.*, 2024c] Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. Harmfully manipulated images matter in multimodal misinformation detection. In *MM*, pages 2262–2271, 2024.

[Wang *et al.*, 2024d] Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, et al. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *ACL*, pages 2351–2374, 2024.

[Wu *et al.*, 2023] Lianwei Wu, Yuan Rao, Cong Zhang, Yongqiang Zhao, and Ambreen Nazir. Category-controlled encoder-decoder for fake news detection. *TKDE*, 35(2):1242–1257, 2023.

[Xiao *et al.*, 2024] Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. Msynfd: Multi-hop syntax aware fake news detection. In *WWW*, pages 4128–4137, 2024.

[Xue *et al.*, 2021] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, pages 483–498, 2021.

[Ying *et al.*, 2023] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. Bootstrapping multi-view representations for fake news detection. In *AAAI*, pages 5384–5392, 2023.

[Yuan *et al.*, 2021] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, pages 27263–27277, 2021.

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

[Zhang *et al.*, 2023] Yongjun Zhang, Hao Lin, Yi Wang, and Xinguang Fan. Sinophobia was popular in chinese language communities on twitter during the early covid-19 pandemic. *Humanities and Social Sciences Communications*, 10(1):1–12, 2023.

[Zhang *et al.*, 2024] Jiajun Zhang, Zhixun Li, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. Evolving to the future: Unseen event adaptive fake news detection on social media. In *CIKM*, pages 4273–4277, 2024.

[Zhu *et al.*, 2022] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. In *SIGIR*, pages 2120–2125, 2022.