Robust Misinformation Detection by Visiting Potential Commonsense Conflict

Bing Wang, Ximing Li, Changchun Li, Bingrui Zhao, Bo Fu, Renchu Guan, Shengsheng Wang*

IJCAI 2025 Guangzhou August 29-31

Motivation

How do human beings identify misinformation?

Recent psychological studies partially offer an answer as human beings naturally distinguish misinformation by referring to their pre-existing **commonsense knowledge**.

> ***In certain scenarios, articles with misinformation are more likely to involve commonsense conflict***

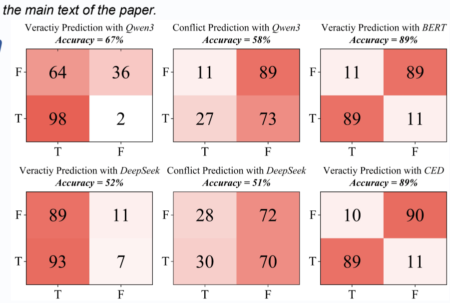
How to measure and express commonsense conflict for given articles?

Large Language Model May be a Bad Choice!

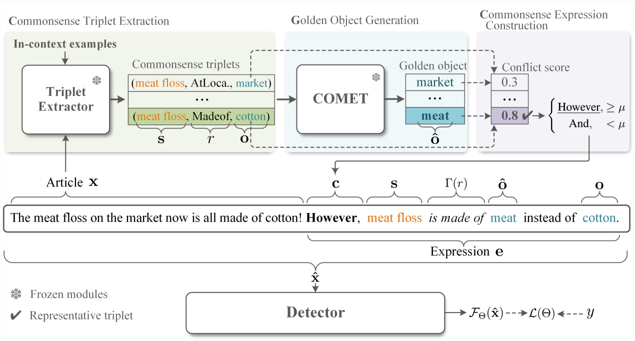
* *This preliminary analysis is not included in the main text of the paper.*

1. LLMs underperform in veracity prediction compared to supervised models

2. LLMs struggle to directly identify commonsense conflicts in articles



Our Proposed Method: MD-PCC



We design a **commonsense template** to express the potential commonsense conflict measured by prevalent commonsense reasoning methods and specify it for each original article as the augmentation.

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

Conjunction | Subject | Relation | Golden Object | Object

However, meat floss is made of meat instead of cotton.

Meat floss on the market now is all made of cotton!

Module 1: Commonsense Triplet Extraction

For each article, we extract a certain number of relevant **commonsense triplets**. To achieve this, we first screen **all relations** to extract all corresponding triplets from the article and then **filter out** the meaningless ones.

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

Module 2: Golden Object Generation

Given subjects and relations, we generate their **golden objects**, which are aligned with real-world commonsense knowledge. Specifically, we feed them into the prevalent **commonsense tool** to generate the golden object.

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

Module 3: Commonsense Expression Construction

We construct a **commonsense expression** by filling the commonsense template. We first compute conflict scores by *BARTScore*. We then select the **highest** conflict score from the set, and use it to fill the commonsense template.

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

A New Dataset: *CoMis*

For Commonsense-Oriented Misinformation Detection

Source	#Num.	fake	real
Weibo-16 [Ma *et al.*, 2016]	523	223	300
Weibo-20 [Zhang *et al.*, 2021]	567	312	255
Weibo-COVID19 [Lin *et al.*, 2022]	69	22	47
Science Facts	313	258	55
Food Rumor	108	77	31
Total	1,580	892	688

❶ **article**: The parasites in sashimi are not terrible, as long as you dip them in wasabi before eating, they can be eliminated.
label: fake **source**: science facts

❷ **article**: Lotus root starch is a powder made from lotus root. It has a unique taste and nutritional value and is a very popular food. Although lotus root starch is good, it should be consumed in moderation to avoid excessive intake.
label: real **source**: science facts

❸ **article**: The New England Journal of Medicine reminds: The remains of beaten mosquito corpses may enter the skin, causing fungal infections and even death!
label: fake **source**: *Weibo-16*

Experimental Results

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**

MD-PCC **outperform five** baselines across **five** datasets

Key Takeaways

- **Motivation:** In certain scenarios, articles with misinformation are more likely to involve commonsense conflict. Meanwhile, large language models may be a bad choice to identify them.
- **Method**: We design a commonsense template to express the potential commonsense conflict measured by prevalent commonsense reasoning methods and specify it for each original article as the augmentation.
- **Experiments**: We construct a new commonsense-oriented dataset *CoMis*. By comparing with baseline models, we have demonstrated the effectiveness of our model.

My homepage

Github Repo