Robust Misinformation Detection by Visiting Potential Commonsense Conflict

Authors: **Bing Wang** (wangbing1416@gmail.com),

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

IJCAI 2025
Guangzhou August 29-31

Code page

My homepage

Background

Social media platforms are full of misinformation, causing lots of damage

Over-the-counter cold and cough medications are being pulled from drugstore shelves in an effort to start the “next plandemic” or force people to get the COVID-19 vaccine.

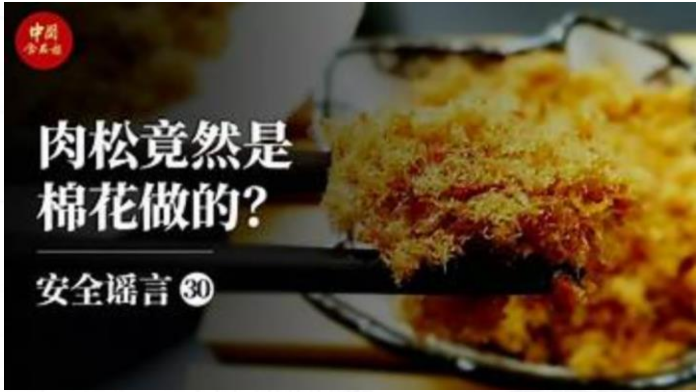
COVID-19 vaccines are safe for people who have existing health conditions, including conditions that have a higher risk of getting serious illness with COVID-19.

Misinformation Detection

Motivation

How do human beings identify misinformation?

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



Meat floss is made of cotton?!

Meat floss is made of ***meat****!*

Meat != Cotton

FAKE

Motivation

How do human beings identify misinformation?

narios, articles with misinformation are

Basic Challenge

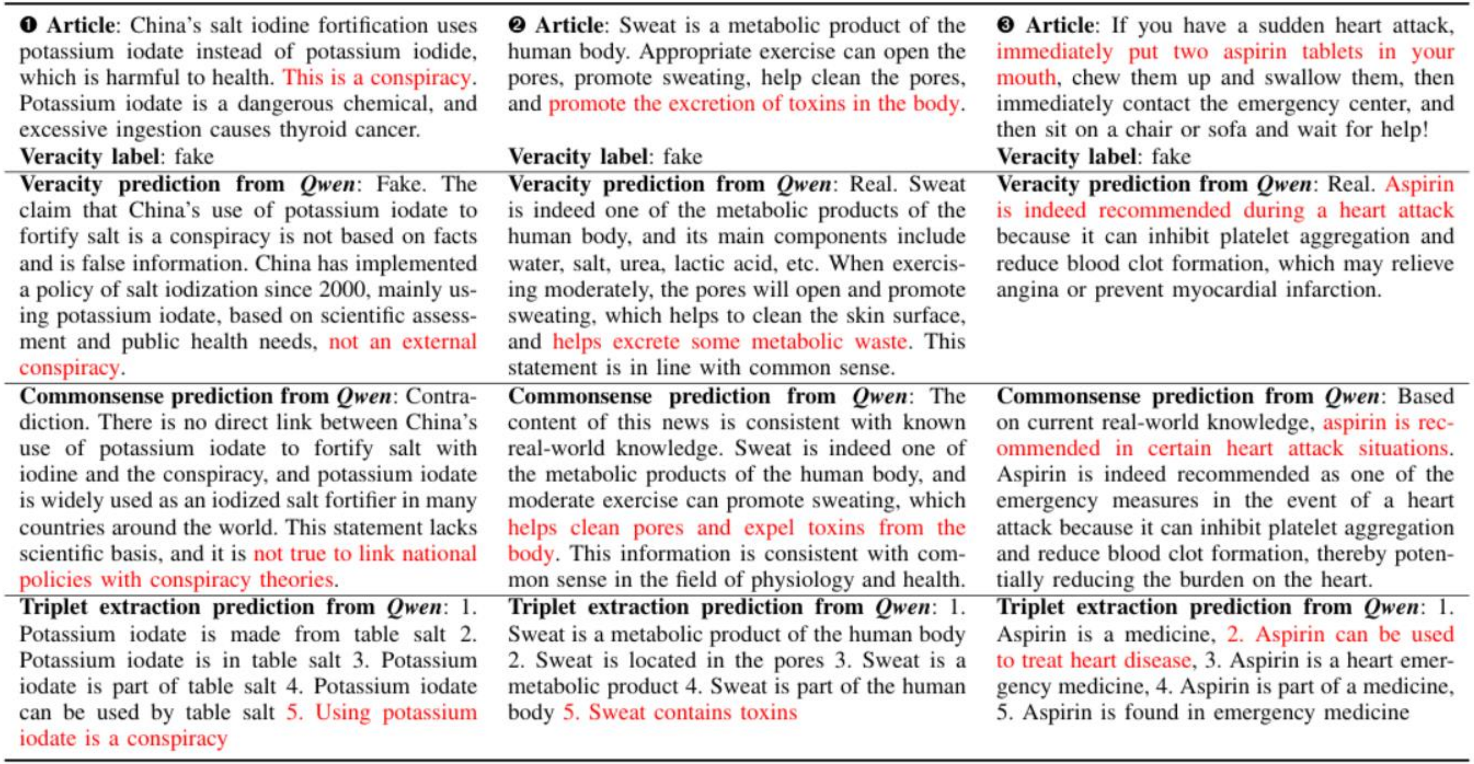
How to measure and express commonsense conflict for given articles?

Meat floss is made of cotton?!

Meat != Cotton

FAKE

Are Large Language Models Good at It?



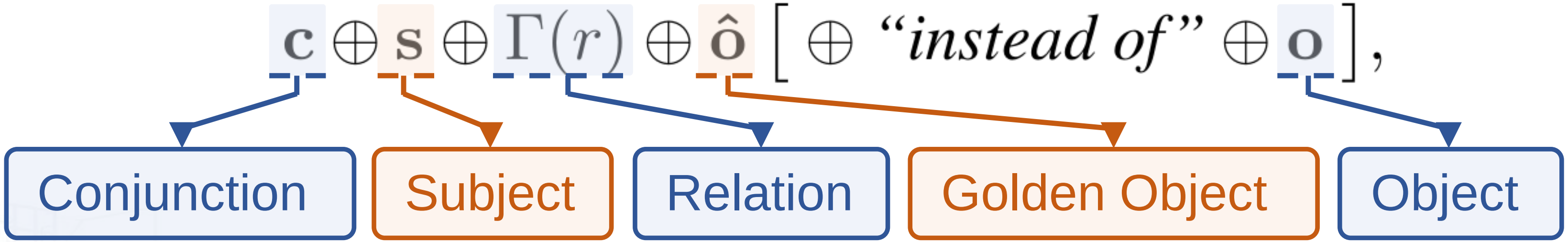
❶ **Article**: China's salt iodine fortification uses potassium iodate instead of potassium iodide, which is harmful to health. This is a conspiracy. Potassium iodate is a dangerous chemical, and excessive ingestion causes thyroid cancer. **Veracity label**: fake	❷ **Article**: Sweat is a metabolic product of the human body. Appropriate exercise can open the pores, promote sweating, help clean the pores, and promote the excretion of toxins in the body. **Veracity label**: fake	❸ **Article**: If you have a sudden heart attack, immediately put two aspirin tablets in your mouth, chew them up and swallow them, then immediately contact the emergency center, and then sit on a chair or sofa and wait for help! **Veracity label**: fake
Veracity prediction from *Qwen*: Fake. The claim that China's use of potassium iodate to fortify salt is a conspiracy is not based on facts and is false information. China has implemented a policy of salt iodization since 2000, mainly using potassium iodate, based on scientific assessment and public health needs, not an external conspiracy.	**Veracity prediction from *Qwen***: Real. Sweat is indeed one of the metabolic products of the human body, and its main components include water, salt, urea, lactic acid, etc. When exercising moderately, the pores will open and promote sweating, which helps to clean the skin surface, and helps excrete some metabolic waste. This statement is in line with common sense.	**Veracity prediction from *Qwen***: Real. Aspirin is indeed recommended during a heart attack because it can inhibit platelet aggregation and reduce blood clot formation, which may relieve angina or prevent myocardial infarction.
Commonsense prediction from *Qwen*: Contradiction. There is no direct link between China's use of potassium iodate to fortify salt with iodine and the conspiracy, and potassium iodate is widely used as an iodized salt fortifier in many countries around the world. This statement lacks scientific basis, and it is not true to link national policies with conspiracy theories.	**Commonsense prediction from *Qwen***: The content of this news is consistent with known real-world knowledge. Sweat is indeed one of the metabolic products of the human body, and moderate exercise can promote sweating, which helps clean pores and expel toxins from the body. This information is consistent with common sense in the field of physiology and health.	**Commonsense prediction from *Qwen***: Based on current real-world knowledge, aspirin is recommended in certain heart attack situations. Aspirin is indeed recommended as one of the emergency measures in the event of a heart attack because it can inhibit platelet aggregation and reduce blood clot formation, thereby potentially reducing the burden on the heart.
Triplet extraction prediction from *Qwen*: 1. Potassium iodate is made from table salt 2. Potassium iodate is in table salt 3. Potassium iodate is part of table salt 4. Potassium iodate can be used by table salt 5. Using potassium iodate is a conspiracy	**Triplet extraction prediction from *Qwen***: 1. Sweat is a metabolic product of the human body 2. Sweat is located in the pores 3. Sweat is a metabolic product 4. Sweat is part of the human body 5. Sweat contains toxins	**Triplet extraction prediction from *Qwen***: 1. Aspirin is a medicine, 2. Aspirin can be used to treat heart disease, 3. Aspirin is a heart emergency medicine, 4. Aspirin is part of a medicine, 5. Aspirin is found in emergency medicine

Actually not, at least on the 7B models

Our Method: MD-PCC

Basic Idea

We design a commonsense template to express the potential commonsense conflict



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

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!

Our Method: MD-PCC

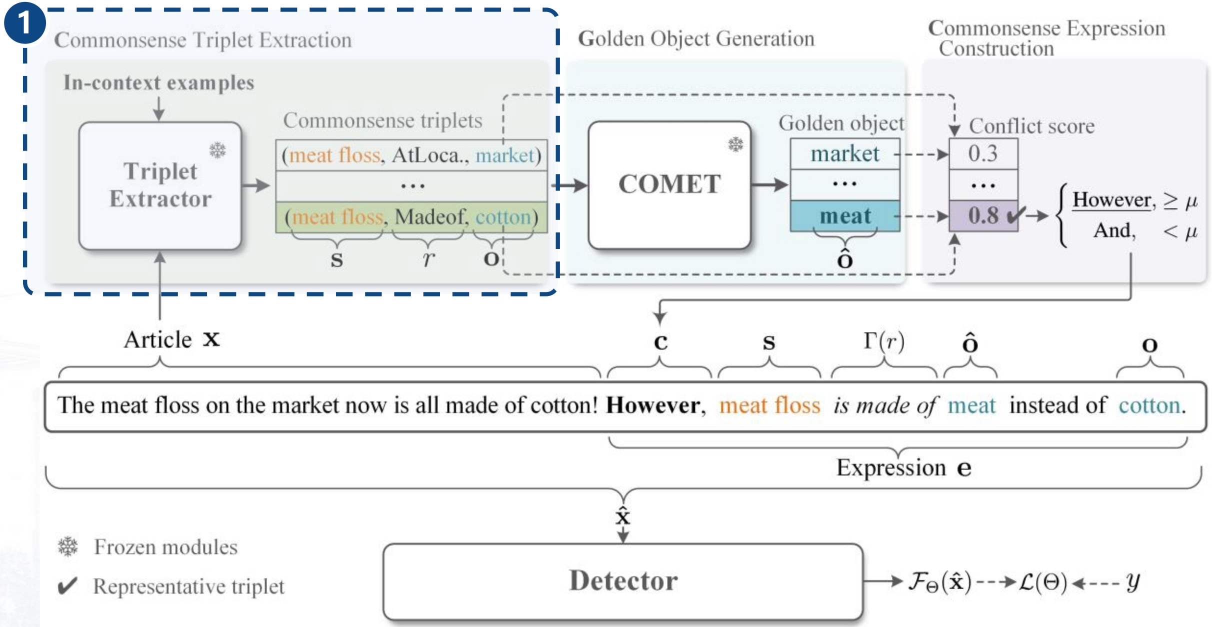
Basic Idea

We design a commonsense template to express the potential commonsense conflict

1 Commonsense Triplet Extraction

# we first screen all relations to extract all relevant triplets from the article.

# then filter the meaningless ones.


1
Commonsense Triplet Extraction
In-context examples
Triplet Extractor
Commonsense triplets
(meat floss, AtLoca., market)
...
(meat floss, Madeof, cotton)
s
r
o
Golden Object Generation
COMET
Golden object
market
...
meat
ô
Commonsense Expression Construction
Conflict score
0.3
...
0.8
However, ≥ μ
And, < μ
Article x
c
s
Γ(r)
ô
o
The meat floss on the market now is all made of cotton! However, meat floss is made of meat instead of cotton.
Expression e
x̂
Frozen modules
✔ Representative triplet
Detector
$\mathcal{F}_\Theta(\hat{\mathbf{x}}) \dashrightarrow \mathcal{L}(\Theta) \dashleftarrow y$

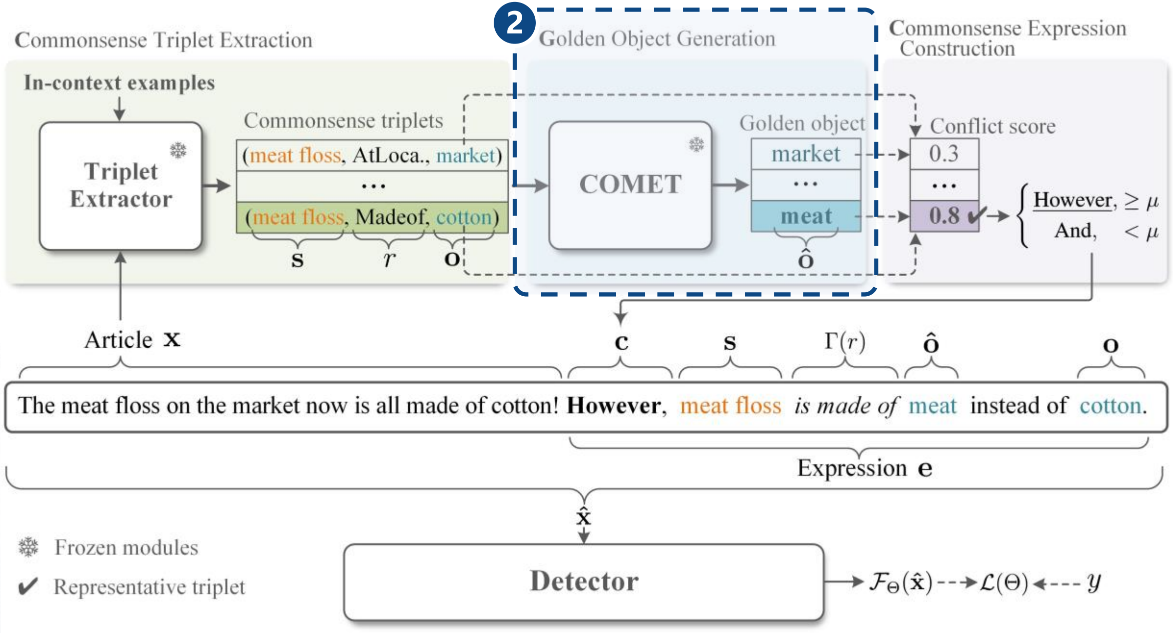
Our Method: MD-PCC

Basic Idea

We design a commonsense template to express the potential commonsense conflict

2 Golden Object Generation

Given subjects and relations, we feed them into the prevalent commonsense tool to generate the golden object.



2
Commonsense Triplet Extraction
In-context examples
Triplet Extractor
Commonsense triplets
(meat floss, AtLoca., market)
...
(meat floss, Madeof, cotton)
s
r
o
Golden Object Generation
COMET
Golden object
market
...
meat
ô
Commonsense Expression Construction
Conflict score
0.3
...
0.8 ✔
However, ≥ μ
And, < μ
Article x
c
s
Γ(r)
ô
o
The meat floss on the market now is all made of cotton! However, meat floss is made of meat instead of cotton.
Expression e
x̂
Detector
$\mathcal{F}_\Theta(\hat{\mathbf{x}}) \dashrightarrow \mathcal{L}(\Theta) \dashleftarrow y$
❄ Frozen modules
✔ Representative triplet

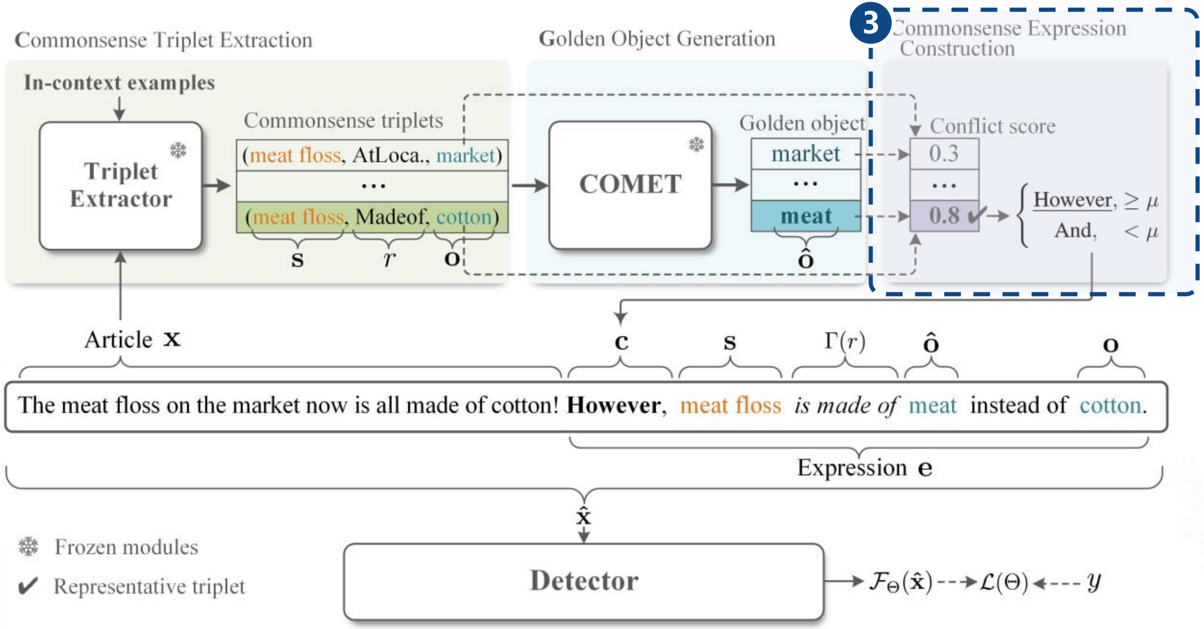
Our Method: MD-PCC

Basic Idea

We design a commonsense template to express the potential commonsense conflict

3 Commonsense Expression Construction

We first compute conflict scores by BARTScore. We then select the highest conflict score, and use it to fill the commonsense template.

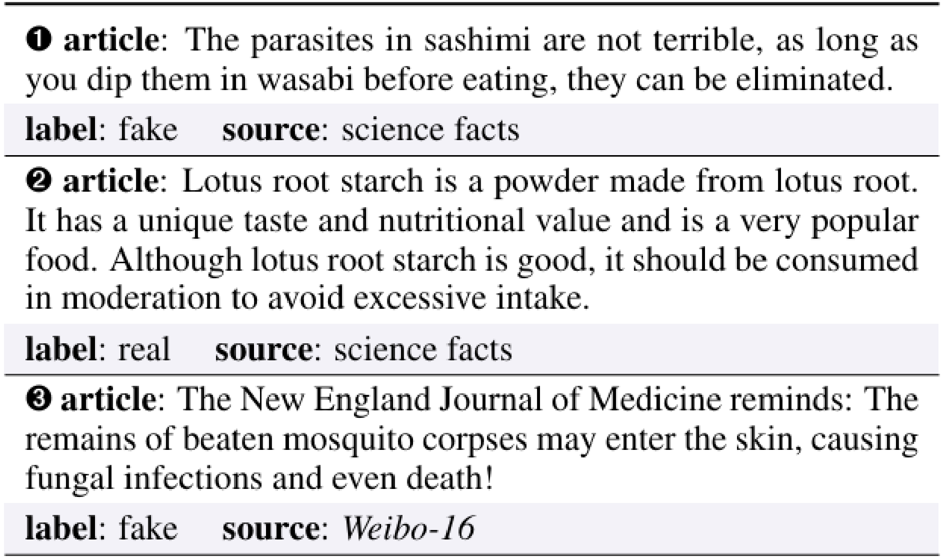


The meat floss on the market now is all made of cotton! **However**, meat floss *is made of* meat instead of cotton.

A New Dataset: CoMis

We collect a new commonsense-oriented misinformation detection dataset: CoMis

Some cases



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

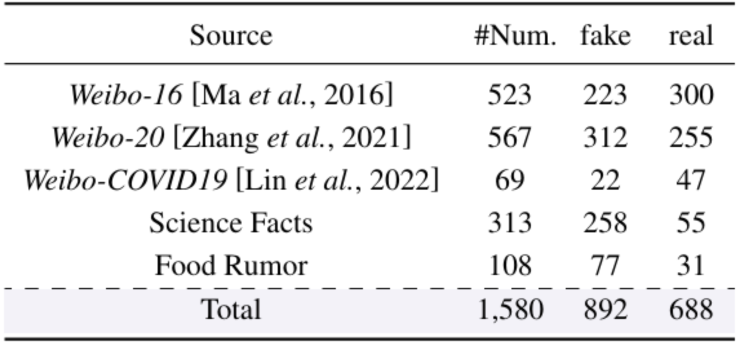
❷ **article**: Lotus root starch is a powder made from lotus root. It has a unique taste and nutritional value and is a very popular food. Although lotus root starch is good, it should be consumed in moderation to avoid excessive intake.

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!

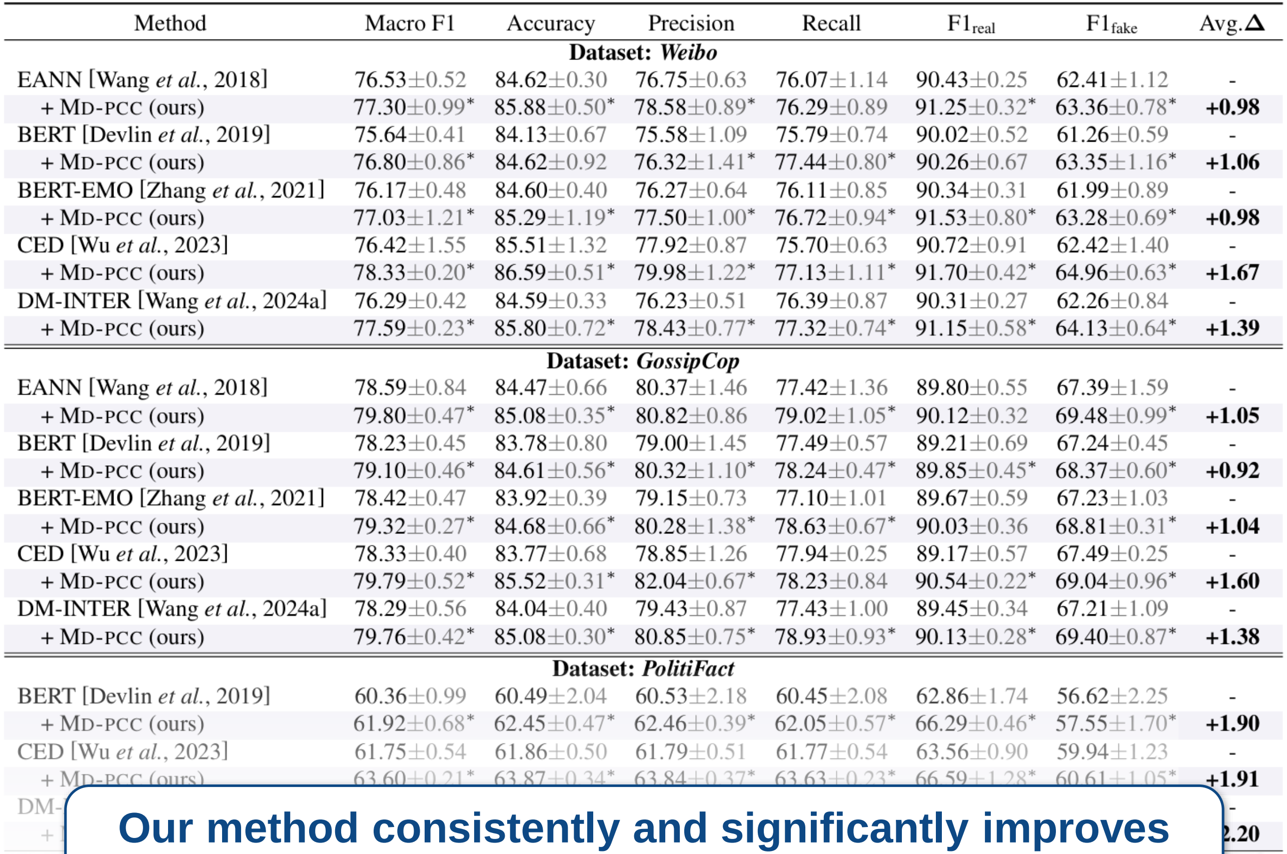
label: fake **source**: *Weibo-16*

Data source



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

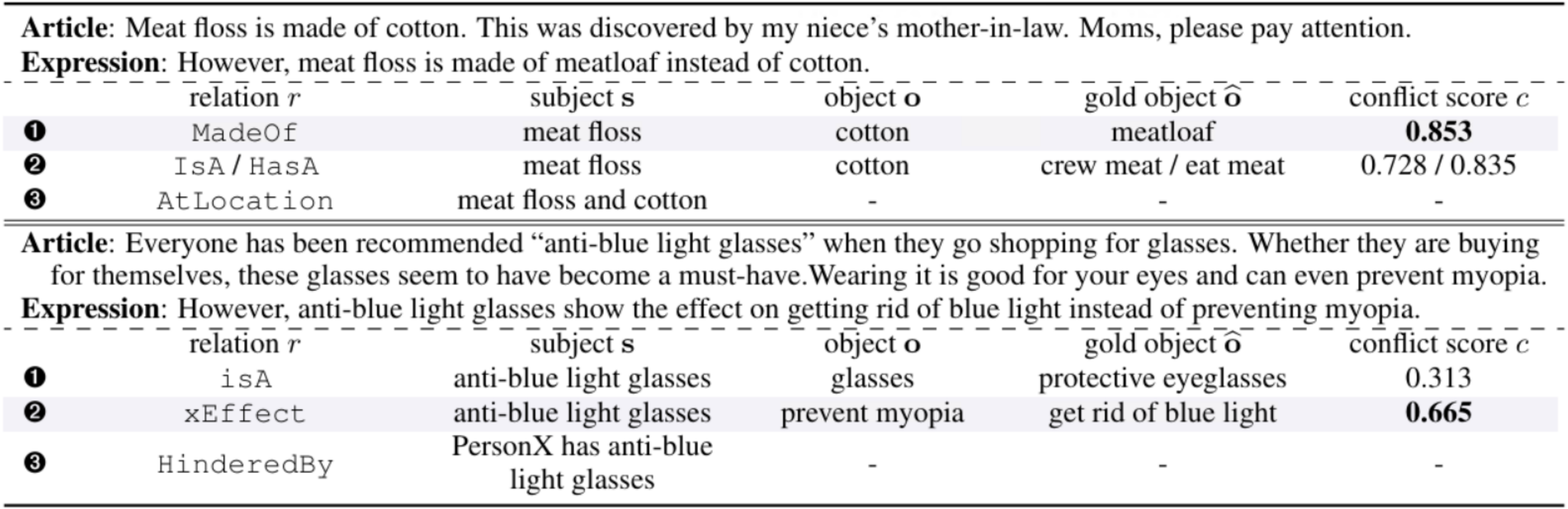
Experimental Results on 5 Baselines and 5 Datasets



Method	Macro F1	Accuracy	Precision	Recall	$F1_{\text{real}}$	$F1_{\text{fake}}$	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-							-
+							**.20**

Our method consistently and significantly improves the performance of baseline models

Case Study



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 $\hat{\mathbf{o}}$	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 $\hat{\mathbf{o}}$	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	-	-	-

Our method gives an accurate identification of commonsense conflicts

Conclusion

- **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.

Thanks.

Robust Misinformation Detection by Visiting Potential Commonsense Conflict

Authors: **Bing Wang** (wangbing1416@gmail.com),

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

IJCAI 2025
Guangzhou August 29-31

Code page

My homepage