LREC-COLING 2024

Towards Robust Evidence-Aware Fake News Detection via Improving Semantic Perception

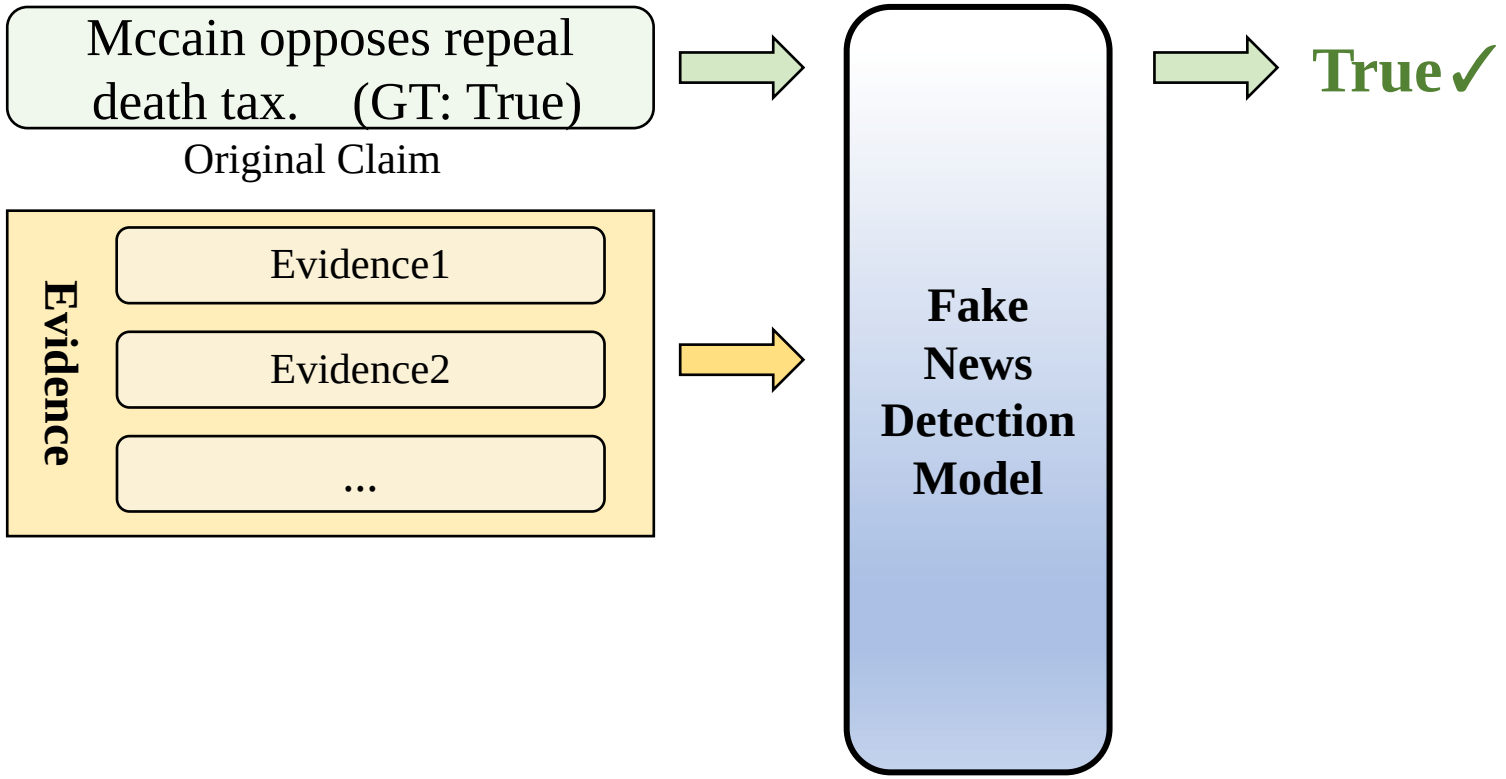
Yike Wu, Yang Xiao, Mengting Hu,
Mengying Liu, Pengcheng Wang, Mingming Liu

Nankai University, Tianjin, China

{wuyike,mthu,liumingming}@nankai.edu.cn

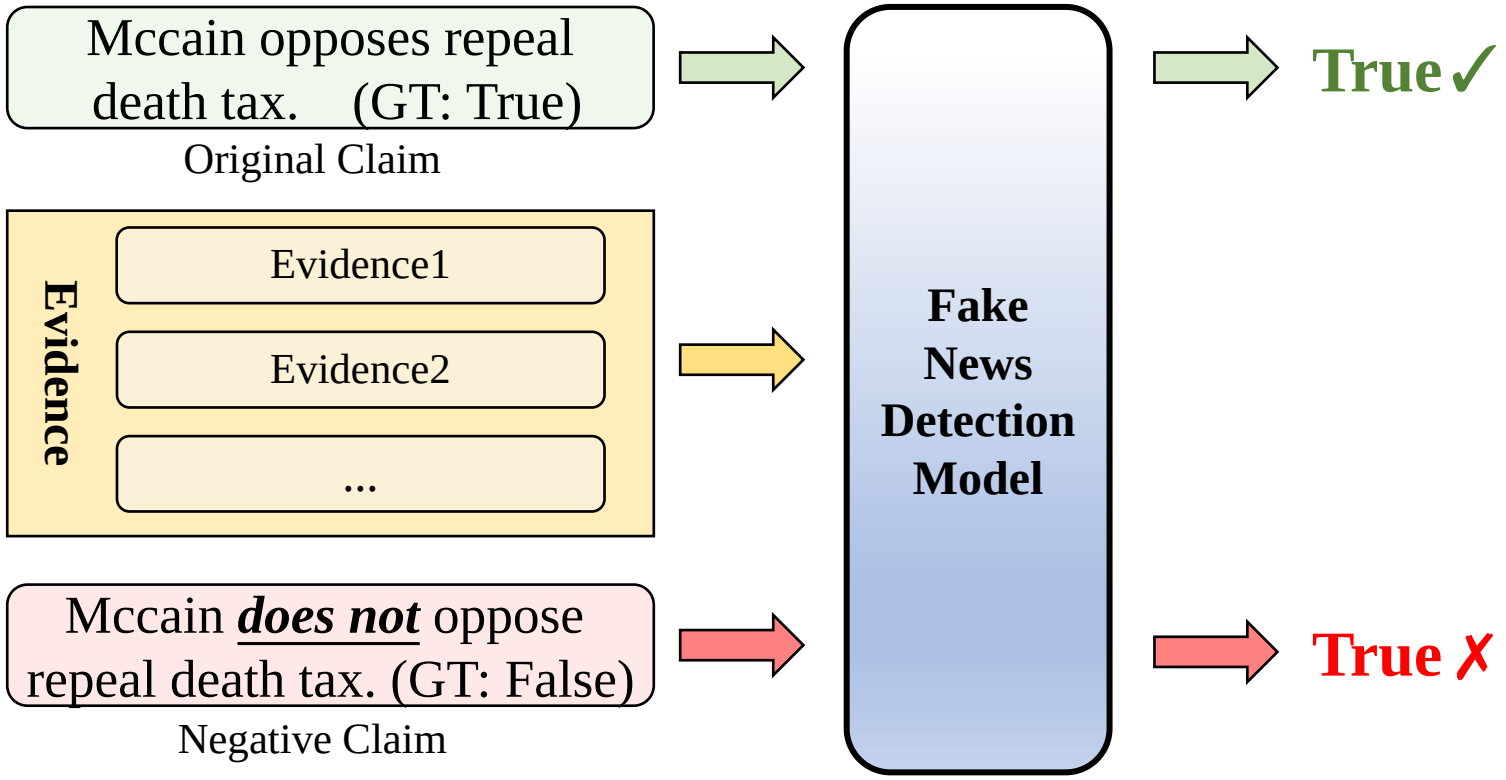
Introduction

- What is evidence-aware fake news detection?



Introduction

- Existing methods lack sufficient awareness to semantics



Pilot Experiment

- A significant decline is observed across evaluation metrics

Dataset	Model	F1-ma	F1-mi	F1-T	F1-F
Snopes	MAC	78.7	83.3	68.7	88.6
	GET	80.0	84.6	70.5	89.5
	BERT	72.8	78.5	60.4	85.2
	RoBERTa	72.2	77.8	59.8	84.7
Snopes -hard	MAC	58.8	60.3	51.2	66.5
	GET	58.7	60.8	49.5	67.9
	BERT	54.4	59.6	39.1	69.7
	RoBERTa	54.2	58.7	40.0	68.5

Original test set

Our extended test set

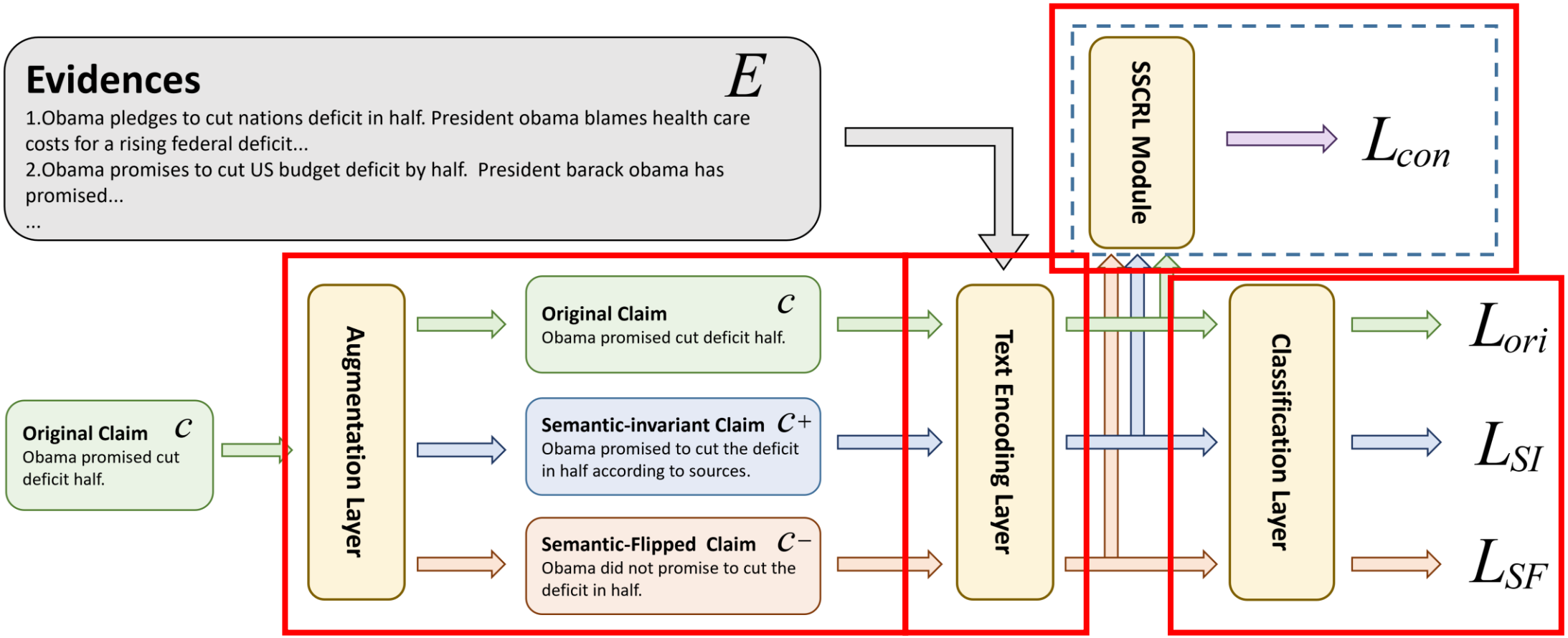
Table 1: Performance (%) of state-of-the-art methods on Snopes and Snopes-hard.

Problem Analysis

- The perspective of dataset
  - The statements in the training set are usually quite different from each other in terms of the semantics and textual expressions

- The perspective of representation learning
  - Existing methods lack explicit encouragement for representation learning to be sensitive to semantic shifts.

Methodology: overview



$$\mathcal{L}_{total} = \mathcal{L}_{ori} + \lambda\mathcal{L}_{con} + \mu\mathcal{L}_{SF} + \nu\mathcal{L}_{SI}$$

Methodology: Semantic-Flipped Augmentation

SpaCy-based

"Mccain opposes repeal death tax"

add negation

"Mccain does not oppose repeal death tax"

ChatGPT-based

Add negation to the following sentences to flip their semantics.
For example.
Mccain opposes repeal death tax.
Mccain does not oppose repeal death tax.

Obtain the semantic-flipped claim c^-

Methodology: Semantic-Invariant Augmentation

PEGASUS-based

"Obama signs bill forgiving student loan debt"

paraphrase

"The student loan debt has been forgiven by Obama through the signing of a bill"

ChatGPT-based

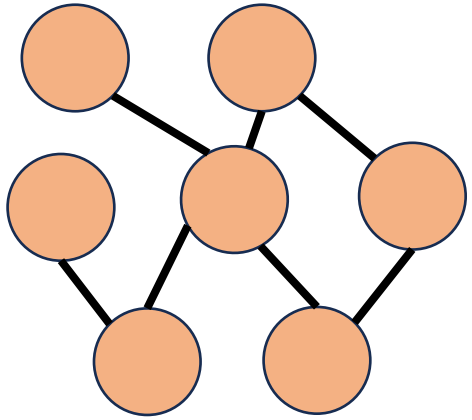
Rewrite each following sentence in a different writing style. Note that you need to keep the original semantics.

Obtain the semantic-invariant claim c^+

Methodology: Text Encoding Layer

Graph-based

Text



PLM-based

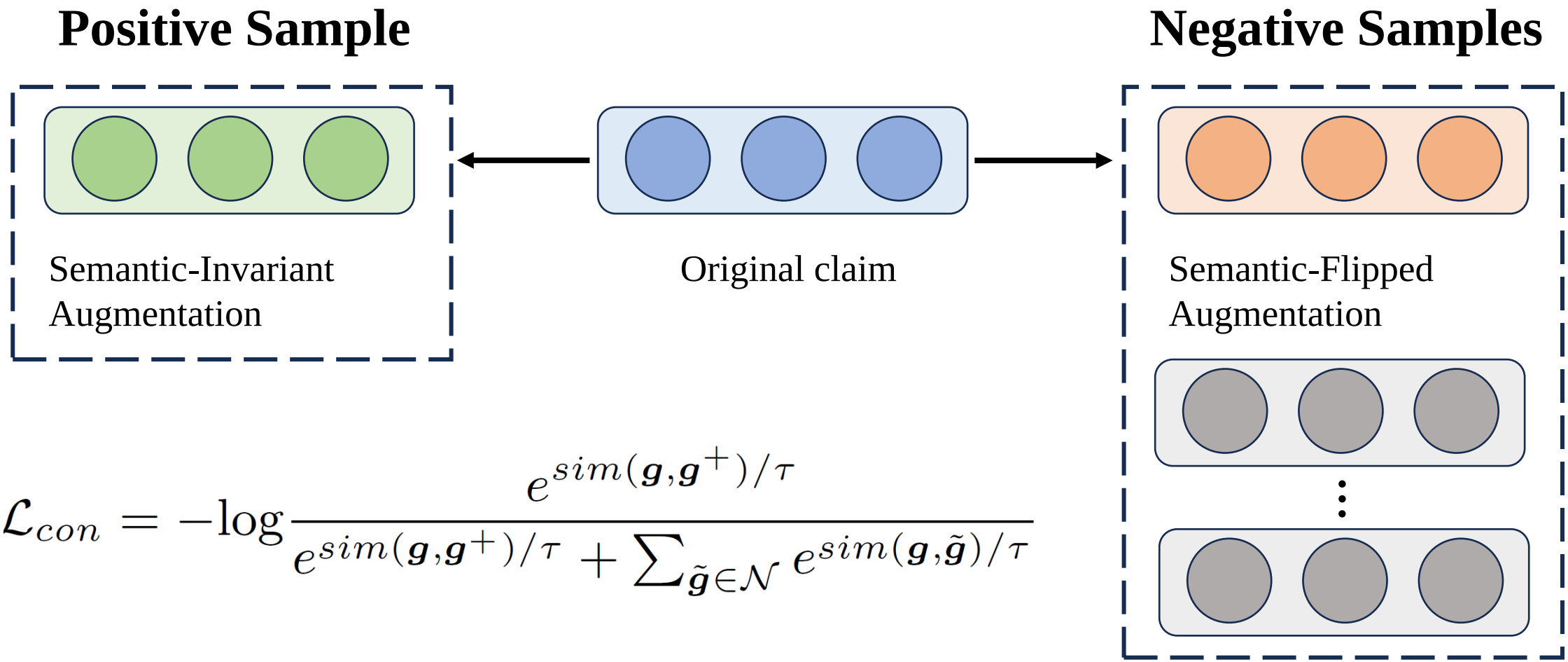
Text

retrieval

most relevant sentences

PLM

Methodology: Semantic-Sensitive Claim Representation Learning Module



$$\mathcal{L}_{con} = -\log \frac{e^{sim(\boldsymbol{g},\boldsymbol{g}^+)/\tau}}{e^{sim(\boldsymbol{g},\boldsymbol{g}^+)/\tau} + \sum_{\tilde{\boldsymbol{g}} \in \mathcal{N}} e^{sim(\boldsymbol{g},\tilde{\boldsymbol{g}})/\tau}}$$

Experiments: Setup

- Datasets
  - Snopes & Politifact
  - Our extended test sets (two versions)
- Evaluation
  - Macro F1、Micro F1
  - F1 score on true/false category
- Baselines
  - Traditional models
  - Pre-trained language models
  - Additional baselines for completeness
    - GET + DA, RoBERTa + DA, ChatGPT

Data Version	Augmentation	Rating Score
conv-hard	Invariant	136.4±17.2
	Flipped	182.7±8.8
gpt-hard	Invariant	174.8±12.6
	Flipped	185.0±8.1

Table 2: Human evaluation on Politifact-hard. "conv-hard" means data generated with conventional approaches, and "gpt-hard" means data generated using ChatGPT. We report the average score with the standard deviation.

Experiments: Comparison with SOTA

Data Version	Model	PolitiFact				Snopes			
		F1-ma	F1-mi	F1-T	F1-F	F1-ma	F1-mi	F1-T	F1-F
original	MAC	68.6	69.1	71.8	65.5	78.7	83.3	68.7	88.6
	GET	69.1	69.4	72.3	66.0	80.0	84.6	70.5	89.5
	BERT	62.8	63.1	64.0	61.5	72.8	78.5	60.4	85.2
	RoBERTa	62.1	65.3	64.6	59.7	72.2	77.8	59.8	84.7
	Ours (conv aug & graph enc)	68.8	69.2	72.2	65.3	76.8	81.9	66.0	87.6
	Ours (gpt aug & graph enc)	68.0	68.3	70.6	65.3	77.2	81.8	66.9	87.5
conv-hard	MAC	56.2	56.2	56.4	56.0	58.1	60.6	48.2	68.0
	GET	56.9	57.1	58.9	54.9	58.1	60.6	48.0	68.2
	BERT	54.8	55.2	54.8	54.9	55.1	59.5	41.1	69.1
	RoBERTa	53.5	54.0	53.5	53.5	54.6	58.9	40.6	68.6
	Ours (conv aug & graph enc)	61.6	61.6	62.5	62.1	77.8	78.3	74.6	81.0
	Ours (conv aug & PLM enc)	**64.3**†	**64.4**†	**65.3**†	**63.4**†	**78.6**†	**79.3**†	**74.7**†	**82.4**†
gpt-hard	MAC	55.4	55.6	57.8	53.0	58.8	60.3	51.2	66.5
	GET	56.5	56.7	58.1	54.9	58.7	60.8	49.5	67.9
	BERT	53.6	54.1	55.2	52.0	54.4	59.6	39.1	69.7
	RoBERTa	53.1	53.8	51.8	54.4	54.2	58.7	40.0	68.5
	GET+DA	56.7	56.7	56.7	56.7	74.0	74.6	69.8	78.1
	RoBERTa+DA	56.7	57.2	54.2	59.3	75.7	76.4	71.9	79.6
	ChatGPT	58.4	58.5	57.8	59.2	58.5	61.5	47.3	69.7
	Ours (gpt aug & graph enc)	61.2	61.2	61.8	60.5	**80.1**†	**80.6**†	**77.0**†	**83.2**†
	Ours (gpt aug & PLM enc)	**62.1**†	**62.2**†	**63.0**†	**61.3**†	77.9	78.5	74.2	81.6

Experiments: Ablation Study

text encoding	#	$\mathcal{L}_{ori}$	$\mathcal{L}_{con}$	$\mathcal{L}_{SF}$	$\mathcal{L}_{SI}$	F1-ma	F1-mi	F1-T	F1-F
graph-based	1	✓				58.7	60.8	49.5	67.9
	2	✓	✓			59.3	61.5	49.8	68.8
	3	✓	✓	✓		71.2	71.8	67.1	75.2
	4	✓	✓	✓	✓	**80.1**	**80.6**	**77.0**	**83.2**
PLM-based	5	✓				54.4	59.6	39.1	69.7
	6	✓	✓			76.9	77.1	74.4	79.3
	7	✓	✓	✓		78.2	78.8	74.6	81.7
	8	✓	✓	✓	✓	**78.6**	**79.4**	**74.8**	**82.5**

Table 4: Ablative performance (%) on the ChatGPT version of Snopes-hard.

Experiments: Model Agnostic Study

Model		PolitiFact				Snopes			
		F1-ma	F1-mi	F1-T	F1-F	F1-ma	F1-mi	F1-T	F1-F
MAC	base	55.4	55.6	57.8	53.0	58.8	60.3	51.2	66.5
	+Ours	59.4 (↑4.0)	59.4(↑3.8)	59.4(↑1.6)	59.4(↑6.4)	78.3(↑19.5)	78.9(↑18.6)	74.8(↑23.6)	81.8(↑15.3)
GET	base	56.5	56.7	58.1	54.9	58.7	60.8	49.5	67.9
	+Ours	61.2(↑4.7)	61.2(↑4.5)	61.8(↑3.7)	60.5(↑5.6)	80.1(↑21.4)	80.6(↑19.8)	77.0(↑27.5)	83.2(↑15.3)
BERT	base	53.6	54.1	55.2	52.0	54.4	59.6	39.1	69.7
	+Ours	61.7(↑8.1)	61.8(↑7.7)	60.8(↑5.6)	62.7(↑10.7)	78.6(↑24.2)	79.4(↑19.8)	74.8(↑35.7)	82.5(↑12.8)
RoBERTa	base	53.1	53.8	51.8	54.4	54.2	58.7	40.0	68.5
	+Ours	62.1(↑9.0)	62.2(↑8.4)	63.0(↑11.2)	61.3(↑6.9)	77.9(↑23.7)	78.5(↑19.8)	74.2(↑34.2)	81.6(↑13.1)

Table 5: Performance (%) of different model architectures on the ChatGPT version of PolitiFact-hard and Snopes-hard. The notation "base" denotes the original method, and "+Ours" denotes applying the proposed framework to the corresponding architecture.

Experiments: Case Study

Claim Version	The Content of the Claim	GET	Ours
Original	Photograph shows bathroom unusually painted floor.	False ✓	False ✓
Semantic-invariant	The photograph depicts a bathroom with an unusually painted floor.	True ✗	False ✓
Semantic-Flipped	Photograph does not show a bathroom with an unusually painted floor.	False ✗	True ✓

(a) An example on models with a graph-based text encoder

Claim Version	The Content of the Claim	RoBERTa	Ours
Original	Obama charge 28 percent tax on home sales.	False ✓	False ✓
Semantic-invariant	Obama imposes a 28% tax on home sales.	True ✗	False ✓
Semantic-Flipped	Obama does not charge 28 percent tax on home sales	False ✗	True ✓

(b) An example on models with a PLM-based text encoder

Conclusion

- A model-agnostic training framework for robust evidence-aware fake news detection
- Two kinds of augmentation to complement the datasets
- Semantic-sensitive claim representation learning module
- Experiments demonstrate the effectiveness of our method

LREC-COLING 2024

Thank you

Yike Wu, Yang Xiao, Mengting Hu,
Mengying Liu, Pengcheng Wang, Mingming Liu

Nankai University, Tianjin, China

{wuyike,mthu,liumingming}@nankai.edu.cn