



# Adversarial Contrastive Learning for Evidence-aware Fake News Detection with Graph Neural Networks

Junfei Wu, Weizhi Xu, Qiang Liu, *Member, IEEE*,  
Shu Wu, *Senior Member, IEEE*, and Liang Wang, *Fellow, IEEE*

TKDE2022

<https://github.com/CRIPAC-DIG/GETRAL>

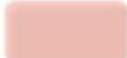
Reported by Xiaoke Li

## Claim

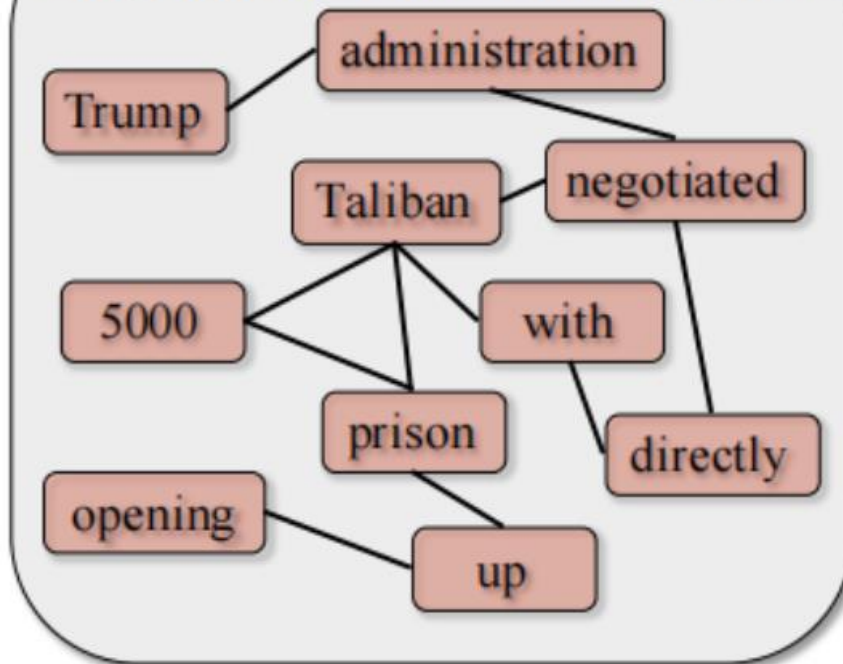
The Trump administration worked to free 5,000 Taliban prisoners.

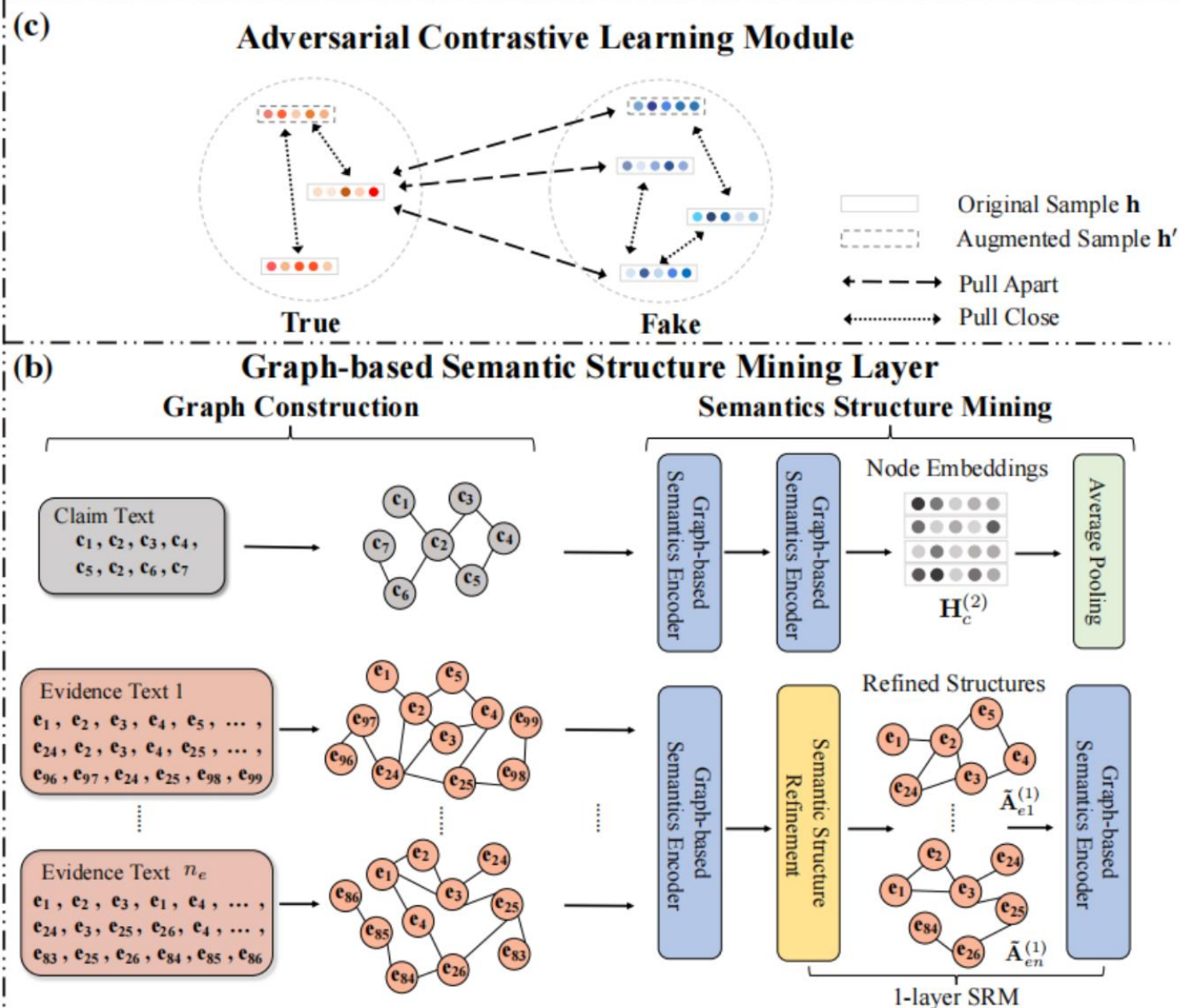
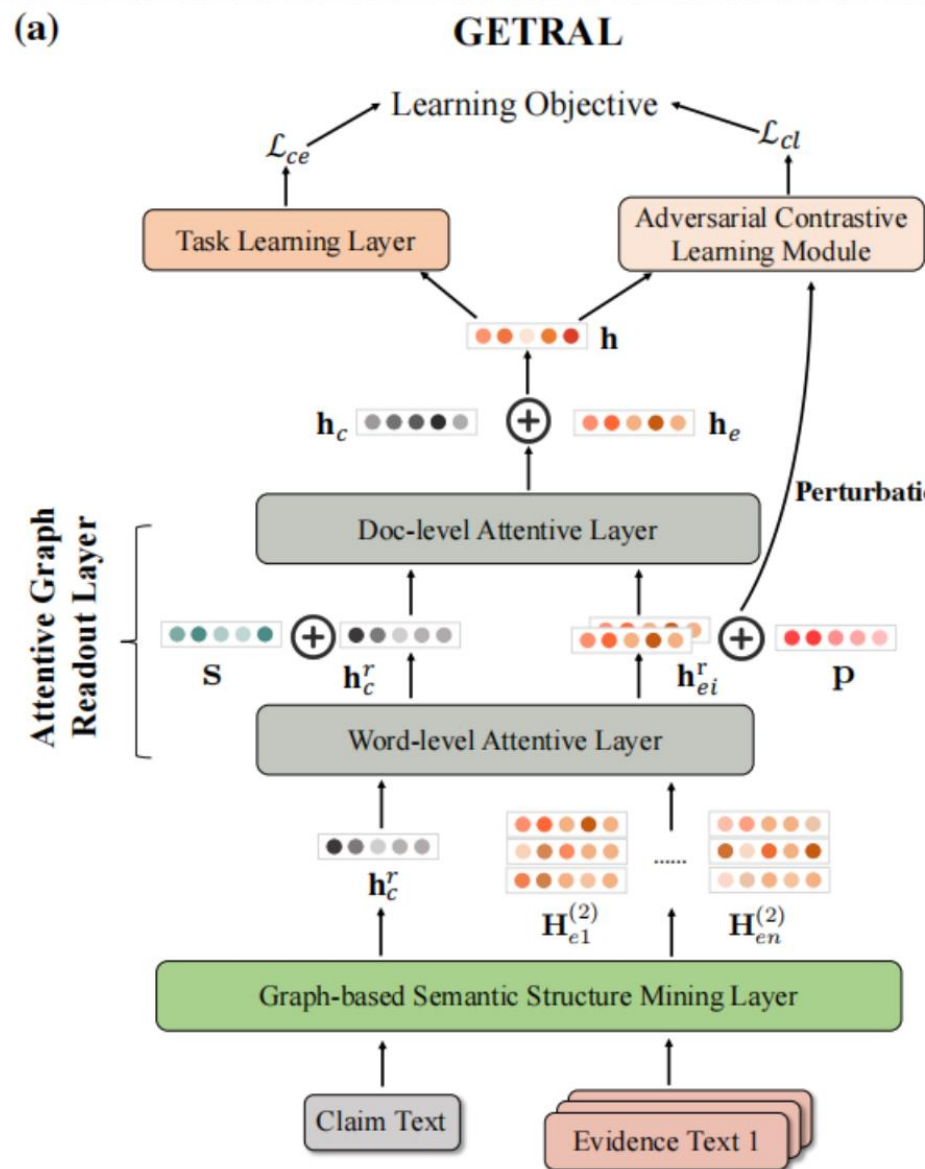
## Evidence

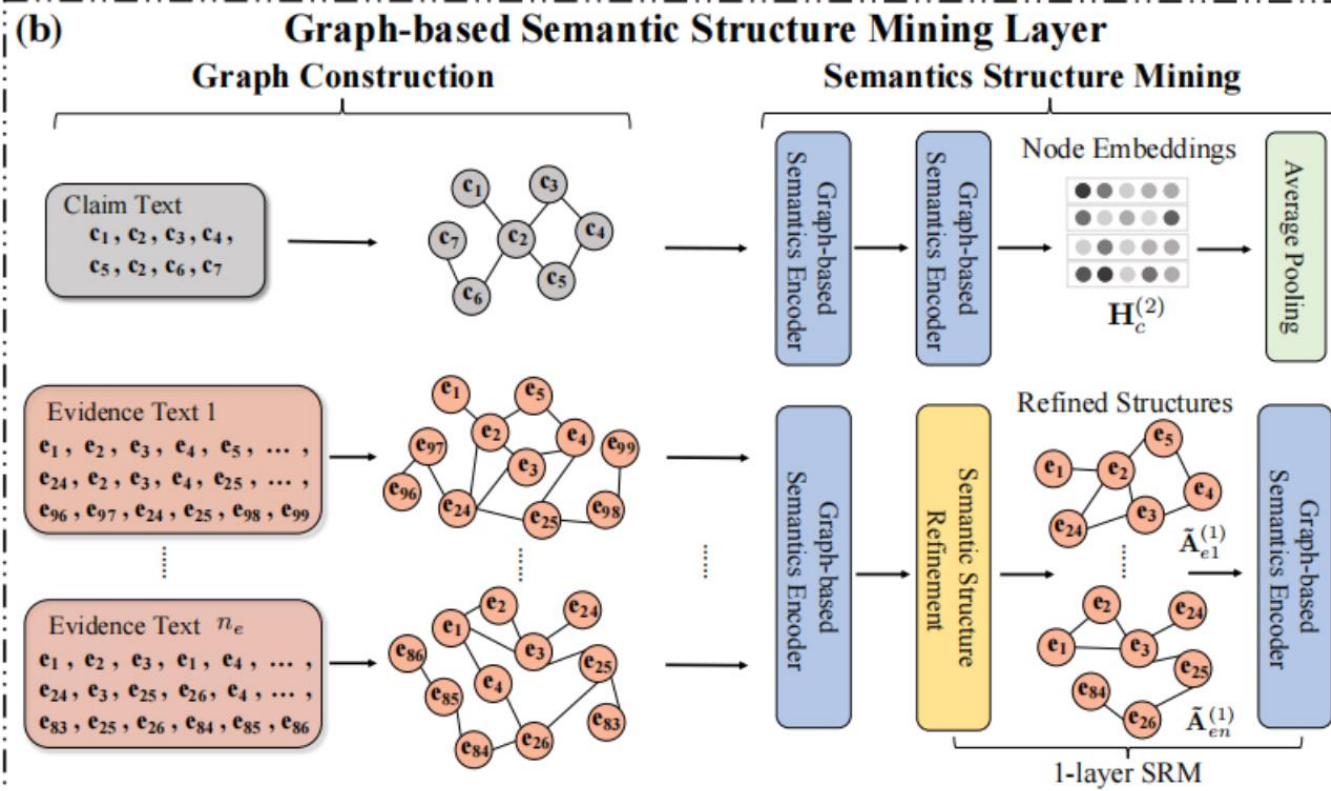
The Trump administration negotiated directly with the Taliban, getting ready to invite them to Camp David, ....., opening up a prison of 5,000 Taliban and probably ISIS-K individuals and letting them free.

 The claim-related snippets

## Subgraph for claim-related snippets







$$\mathbf{a}_i = \sum_{(w_i, w_j) \in \mathcal{C}} \tilde{\mathbf{A}}_{ij} \mathbf{W}_a \mathbf{H}_j \quad (1)$$

$$\mathbf{z}_i = \sigma(\mathbf{W}_z \mathbf{a}_i + \mathbf{U}_z \mathbf{H}_i + \mathbf{b}_z) \quad (2)$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r \mathbf{a}_i + \mathbf{U}_r \mathbf{H}_i + \mathbf{b}_r) \quad (3)$$

$$\tilde{\mathbf{H}}_i = \tanh(\mathbf{W}_h \mathbf{a}_i + \mathbf{U}_h (\mathbf{r}_i \odot \mathbf{H}_i) + \mathbf{b}_h) \quad (4)$$

$$\hat{\mathbf{H}}_i = \tilde{\mathbf{H}}_i \odot \mathbf{z}_i + \mathbf{H}_i \odot (1 - \mathbf{z}_i) \quad (5)$$

$$\mathbf{S}_{se} = \hat{\mathbf{H}}_e \mathbf{W}_{se} \quad (6)$$

$$\mathbf{K}_i = [\mathbf{K}_{i1}; \mathbf{K}_{i2}; \dots; \mathbf{K}_{ik}] \quad (7)$$

$$\mathbf{K}_{it} = \log \sum_j \exp\left(-\frac{(\mathbf{M}_{ij} - \mu_t)^2}{2\sigma_t^2}\right) \quad (8)$$

$$\mathbf{S}_{sc} = \mathbf{K} \mathbf{W}_{sc} \quad (9)$$

$$\mathbf{M}_{ij} = \cos(\mathbf{H}_{ei}, \mathbf{H}_{cj}).$$

$$\mathbf{s}_{se} = \text{GGNN}(\tilde{\mathbf{A}}, \mathbf{S}_{sc}) \quad (10)$$

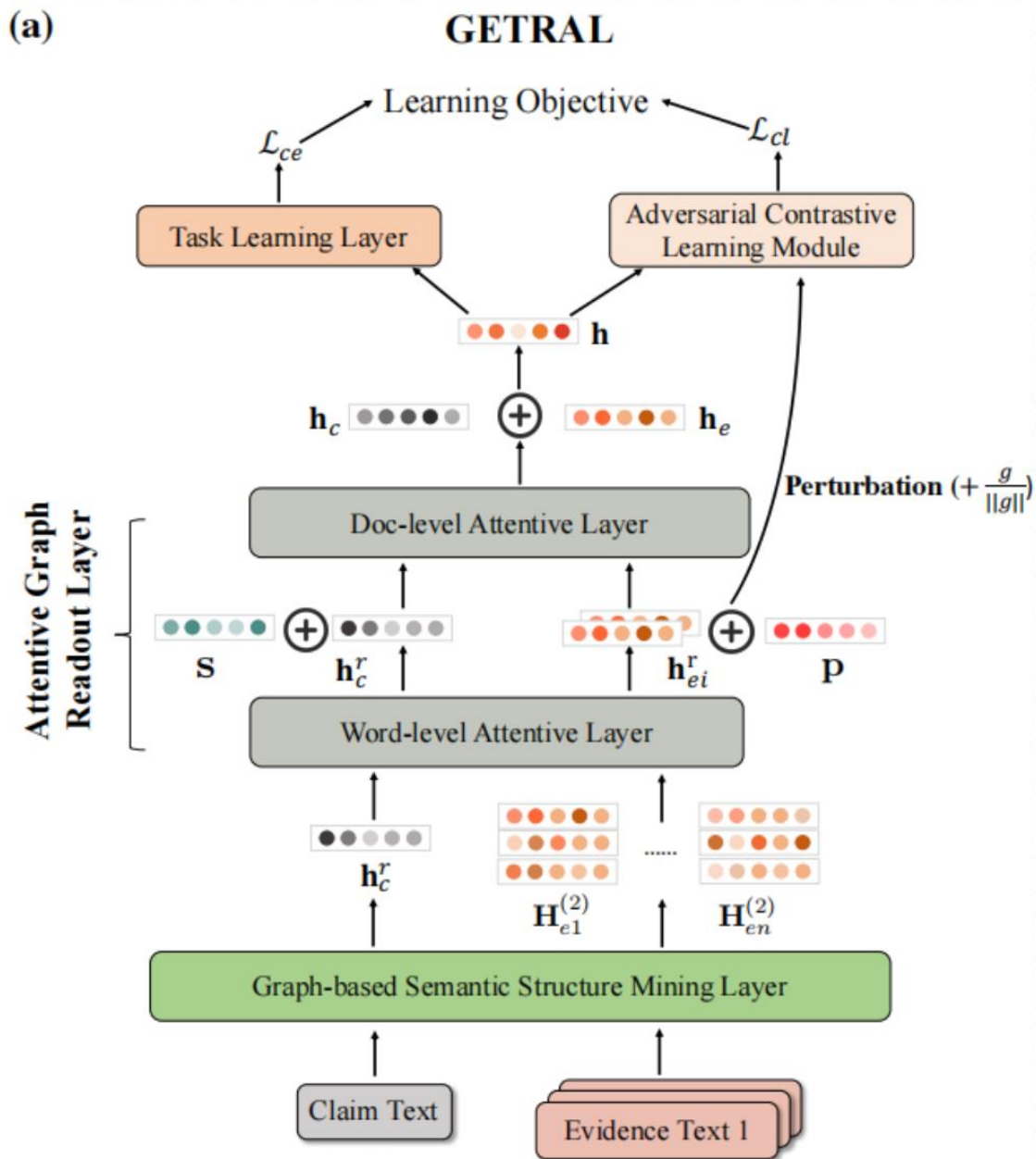
$$\mathbf{s}_{sc} = \text{GGNN}(\tilde{\mathbf{A}}, \mathbf{S}_{se}) \quad (11)$$

$$\mathbf{s}_r = (1 - \beta) \mathbf{s}_{se} + \beta \mathbf{s}_{sc} \quad (12)$$

$$idx = \text{topk\_index}(\mathbf{s}_r) \quad (13)$$

$$\tilde{\mathbf{A}}_{idx, :} = \tilde{\mathbf{A}}_{:, idx} = 0 \quad (14)$$

$$\mathbf{a}_i = \sum_{(w_i, w_j) \in \mathcal{C}} \tilde{\mathbf{A}}_{ij} \mathbf{W}_a \mathbf{H}_j (1 - \sigma(\mathbf{s}_{rj})) \quad (15)$$



$$\mathbf{h}_c^r = \frac{1}{l_c} \sum_{i=1}^{l_c} \mathbf{H}_{ci} \quad (16)$$

$$\mathbf{p}_j = \tanh([\mathbf{H}_{ej}; \mathbf{h}_c^r] \mathbf{W}_c) \quad (17)$$

$$\alpha_j = \frac{\exp(\mathbf{p}_j \mathbf{W}_p)}{\sum_{i=1}^{l_e} \exp(\mathbf{p}_i \mathbf{W}_p)} \quad (18)$$

$$\mathbf{h}_e^r = \sum_{j=1}^{l_e} \alpha_j \mathbf{H}_{ej} \quad (19)$$

$$\mathbf{h}_c = [\mathbf{h}_c^r; \mathbf{s}] \quad (20)$$

$$\mathbf{h}_e^g = [\mathbf{h}_e^r; \mathbf{p}] \quad (21)$$

$$\mathbf{H}_e^g = [\mathbf{h}_{e1}^g; \mathbf{h}_{e2}^g; \dots; \mathbf{h}_{en}^g] \quad (22)$$

$$\mathbf{h}_e = \text{ATTN}(\mathbf{H}_e^g, \mathbf{h}_c) \quad (23)$$

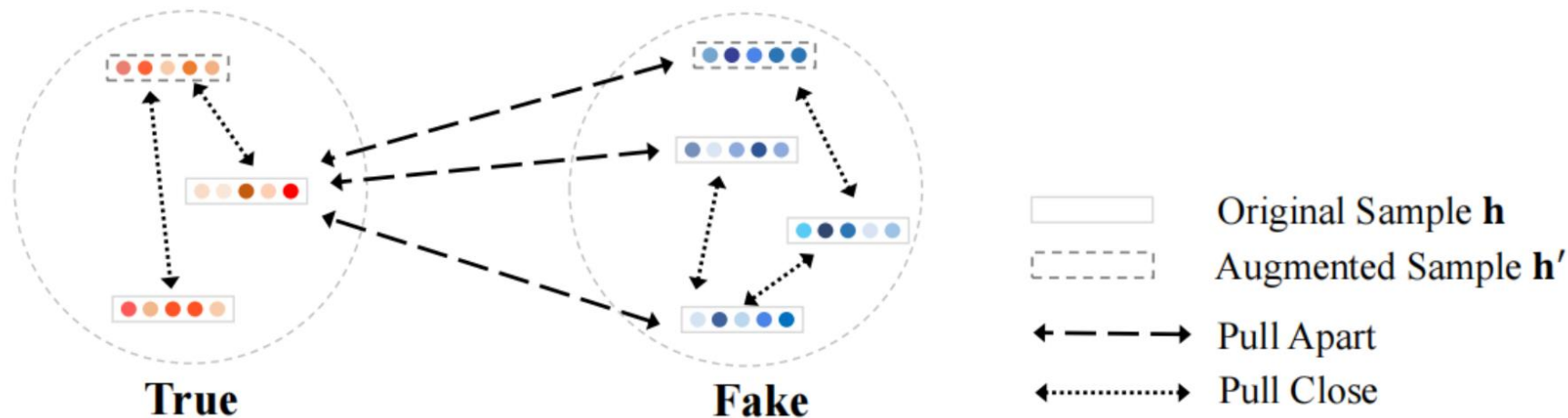
$$\mathbf{h} = [\mathbf{h}_c; \mathbf{h}_e] \quad (24)$$

$$\hat{y} = \text{Softmax}(\mathbf{W}_f \mathbf{h} + \mathbf{b}_f) \quad (25)$$

$$\mathcal{L}_{ce} = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \quad (26)$$

(c)

## Adversarial Contrastive Learning Module



$$\mathcal{L}_{cl} = \frac{-1}{|\mathcal{P}(\mathbf{h})|} \sum_{\mathbf{h}_p \in \mathcal{P}(\mathbf{h})} \log \frac{\exp(\cos(\mathbf{h}, \mathbf{h}_p)/\tau)}{\sum_{\mathbf{h}_n \in \mathcal{N}(\mathbf{h})} \exp(\cos(\mathbf{h}, \mathbf{h}_n)/\tau)} \quad (27)$$

$$\mathbf{g}_{ek} = \nabla_{\mathbf{h}_{ek}^g} \mathcal{L}_{ce} \quad (28)$$

$$\mathbf{h}_{ek}^{g'} = \mathbf{h}_{ek}^g + \epsilon \frac{\mathbf{g}_{ek}}{\|\mathbf{g}_{ek}\|} \quad (29)$$

$$\mathbf{H}_e^{g'} = [\mathbf{h}_{e1}^g; \dots; \mathbf{h}_{ek}^{g'}; \dots] \quad (30)$$

$$\mathbf{h}'_e = \text{ATTN}(\mathbf{H}_e^{g'}, \mathbf{h}_c) \quad (31)$$

$$\mathbf{h}' = [\mathbf{h}_c; \mathbf{h}'_e] \quad (32)$$

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cl} \quad (33)$$



TABLE 1

The statistics of two datasets. The symbol “#” denotes “the number of”. “True” and “False” stand for true claims and false claims, respectively. “Evi.,” “Spe.,” and “Pub.” denote evidences, speakers and publishers.

Dataset	# True	# False	# Evi.	# Spe.	# Pub.
Snopes	1164	3177	29242	N/A	12236
PolitiFact	1867	1701	29556	664	4542



Method	Snopes							
	F1-Ma	F1-Mi	F1-T	P-T	R-T	F1-F	P-F	R-F
LSTM	62.10	71.87	42.95	48.42	39.69	81.25	79.14	83.67
TextCNN	63.08	72.01	45.00	48.16	43.04	81.16	79.88	82.62
BERT	62.05	71.62	43.07	47.73	40.65	81.04	79.31	82.97
DeClarE	72.54	78.61	59.43	61.03	57.93	85.67	85.25	86.39
HAN	75.21	80.23	63.58	62.50	64.69	86.83	87.64	86.11
EHIAN	78.43	82.83	68.41	61.69	76.79	88.47	88.18	89.04
MAC	78.66	83.32	68.74	70.00	68.60	88.58	88.62	88.71
CICD	78.92	83.73	69.07	63.20	<b>77.48</b>	89.30	<b>88.99</b>	89.54
GETRAL	<b>80.61<sup>‡</sup></b>	<b>85.12<sup>‡</sup></b>	<b>71.26<sup>‡</sup></b>	<b>74.18<sup>‡</sup></b>	68.79	<b>89.96<sup>‡</sup></b>	88.90	<b>91.04<sup>‡</sup></b>

PolitiFact							
F1-Ma	F1-Mi	F1-T	P-T	R-T	F1-F	P-F	R-F
60.56	60.87	61.82	63.19	61.27	59.31	59.05	60.43
60.38	60.74	61.52	63.01	61.03	59.24	59.05	60.42
59.71	59.81	60.81	61.95	59.90	58.60	57.73	59.70
65.31	65.25	67.49	66.71	68.32	63.11	63.70	62.46
66.12	66.01	67.92	67.58	68.20	64.33	64.97	63.73
67.22	67.95	68.92	68.64	69.34	65.52	67.49	63.60
68.03	68.25	71.78	67.54	73.49	64.28	67.61	61.68
68.18	68.48	70.24	68.92	71.44	65.72	69.12	62.93
<b>69.53<sup>‡</sup></b>	<b>69.81<sup>‡</sup></b>	<b>72.21<sup>‡</sup></b>	<b>69.73<sup>‡</sup></b>	<b>75.10<sup>‡</sup></b>	<b>66.84<sup>‡</sup></b>	<b>70.26<sup>‡</sup></b>	<b>64.01<sup>‡</sup></b>



TABLE 3

The performance comparison between GETRAL and model variants.

Method	Snopes		PolitiFact	
	F1-Ma	F1-Mi	F1-Ma	F1-Mi
GETRAL-SE-CL	77.51	82.31	67.47	67.77
GETRAL-GSE-CL	78.66	83.32	68.03	68.25
GETRAL-SSR-CL	79.49	84.10	68.45	68.78
GETRAL-CL	80.12	84.52	69.25	69.60
GETRAL-AD	80.32	84.81	69.40	69.69
GETRAL	80.61	85.12	69.53	69.81

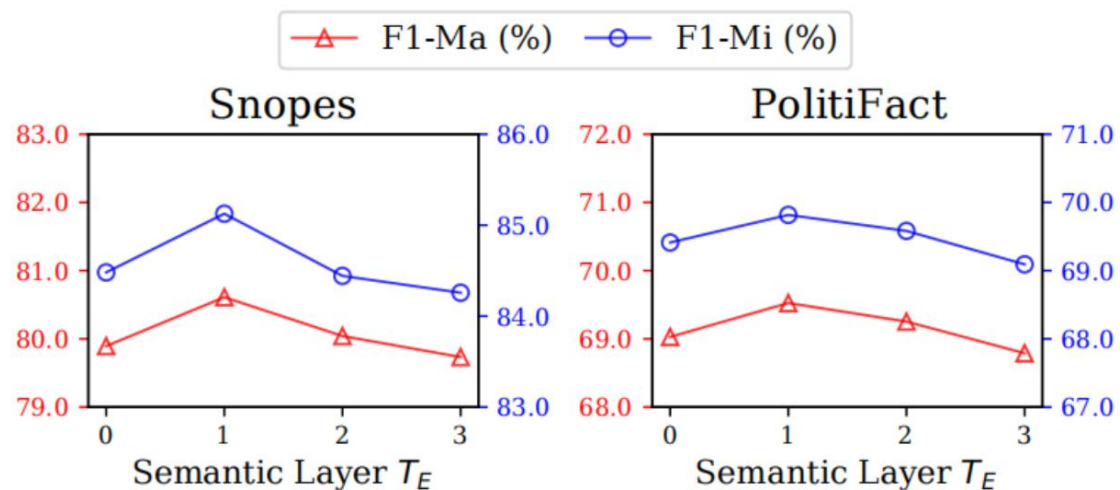


Fig. 3. The influence of different semantics encoder layers  $T_E$  for claims on model performance.

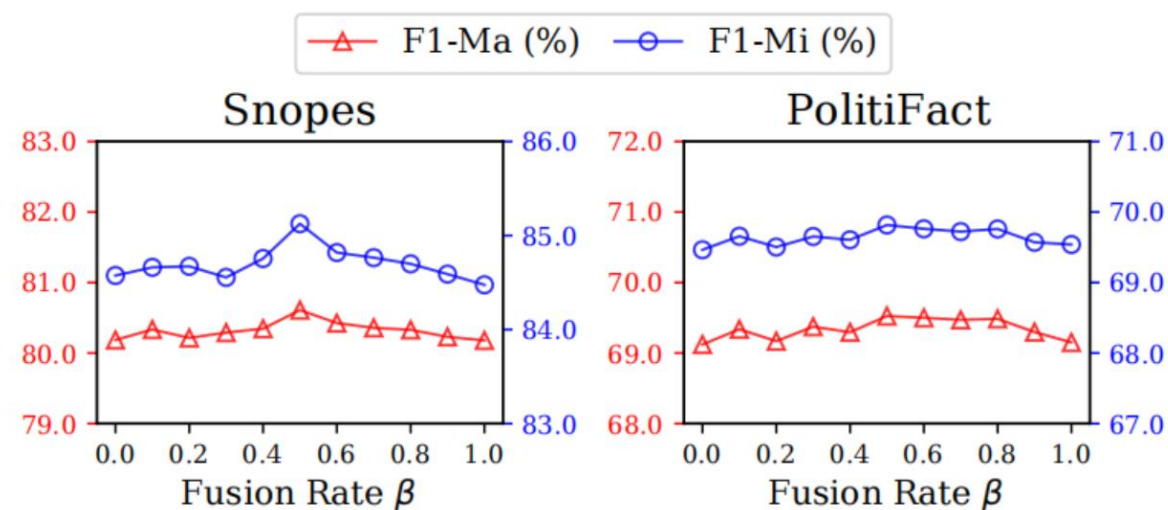


Fig. 4. The influence of different fusion rate  $\beta$  on model performance.

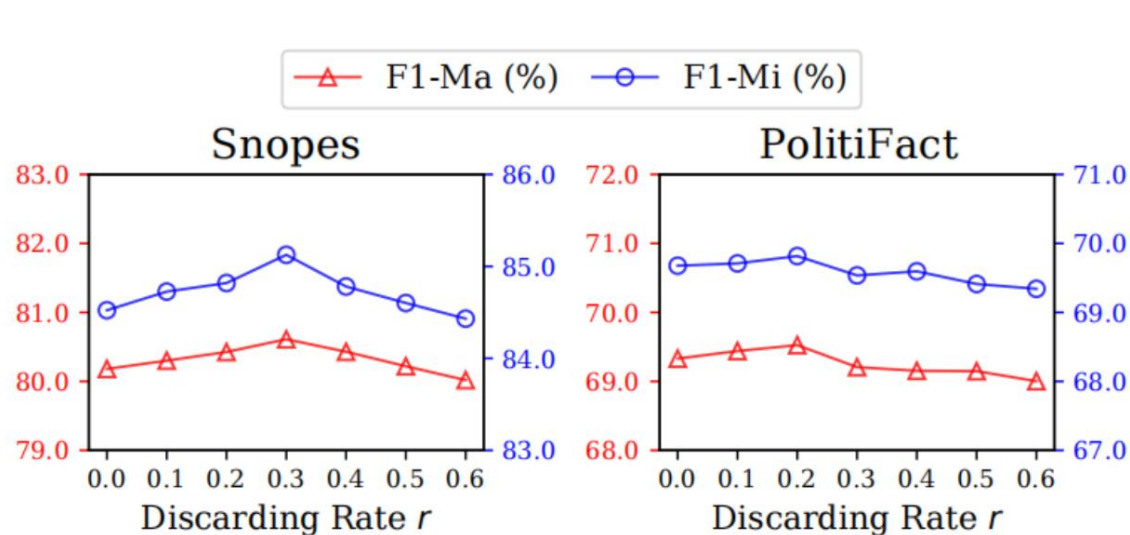


Fig. 5. The influence of different discarding rates  $r$  on model performance.

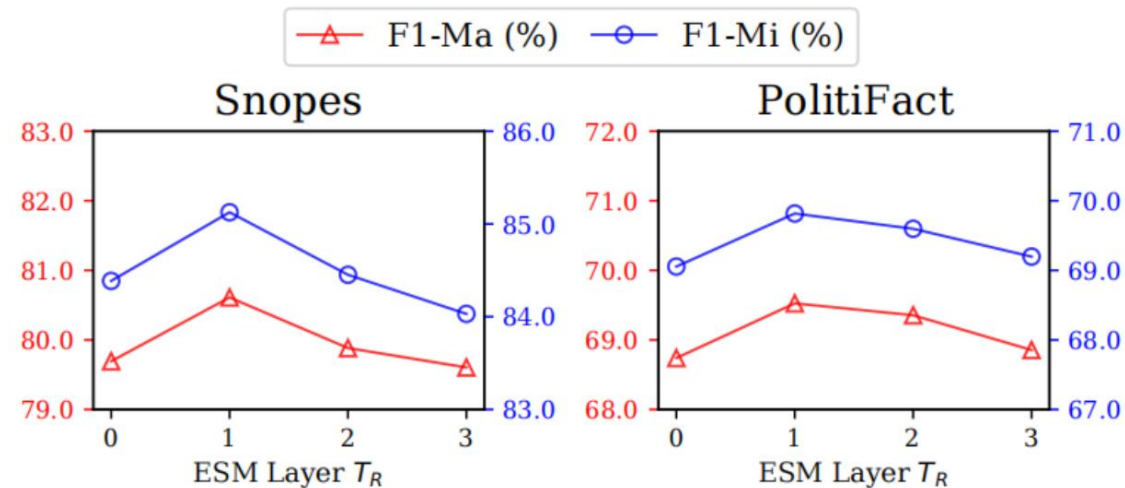


Fig. 6. The influence of different semantic refinement and miner layers  $T_R$  on model performance.

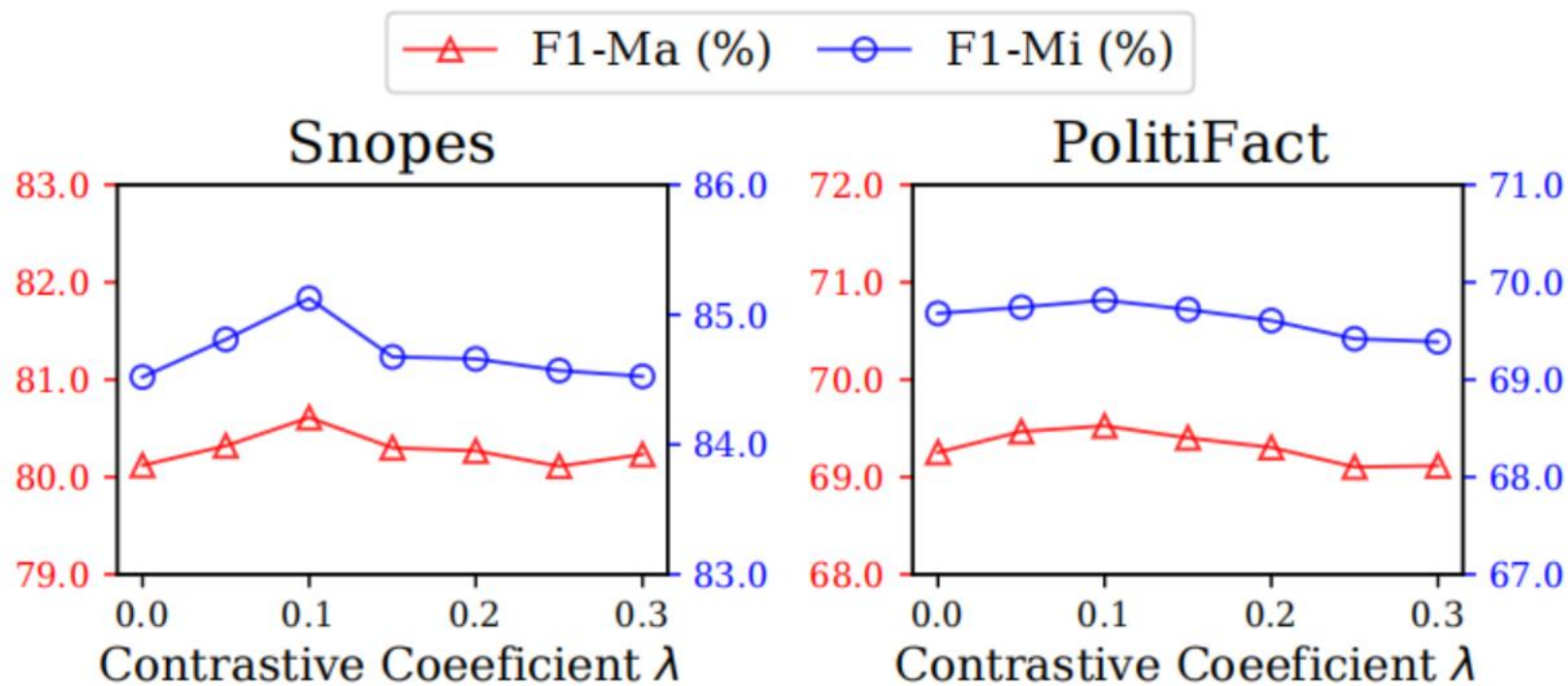


Fig. 7. The influence of different contrastive coefficient  $\lambda$  on model performance.



# Thanks