End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models

Barry Menglong Yao
University at Buffalo
myao2@buffalo.edu

Aditya Shah
Virginia Tech
aditya31@vt.edu

Lichao Sun
Lehigh University
lis221@lehigh.edu

Jin-Hee Cho
Virginia Tech
jicho@vt.edu

Lifu Huang
Virginia Tech
lifuh@vt.edu

SIGIR2023
code:https://github.com/VT-NLP/Mocheg

Reported by Xiaoke Li

Introduction

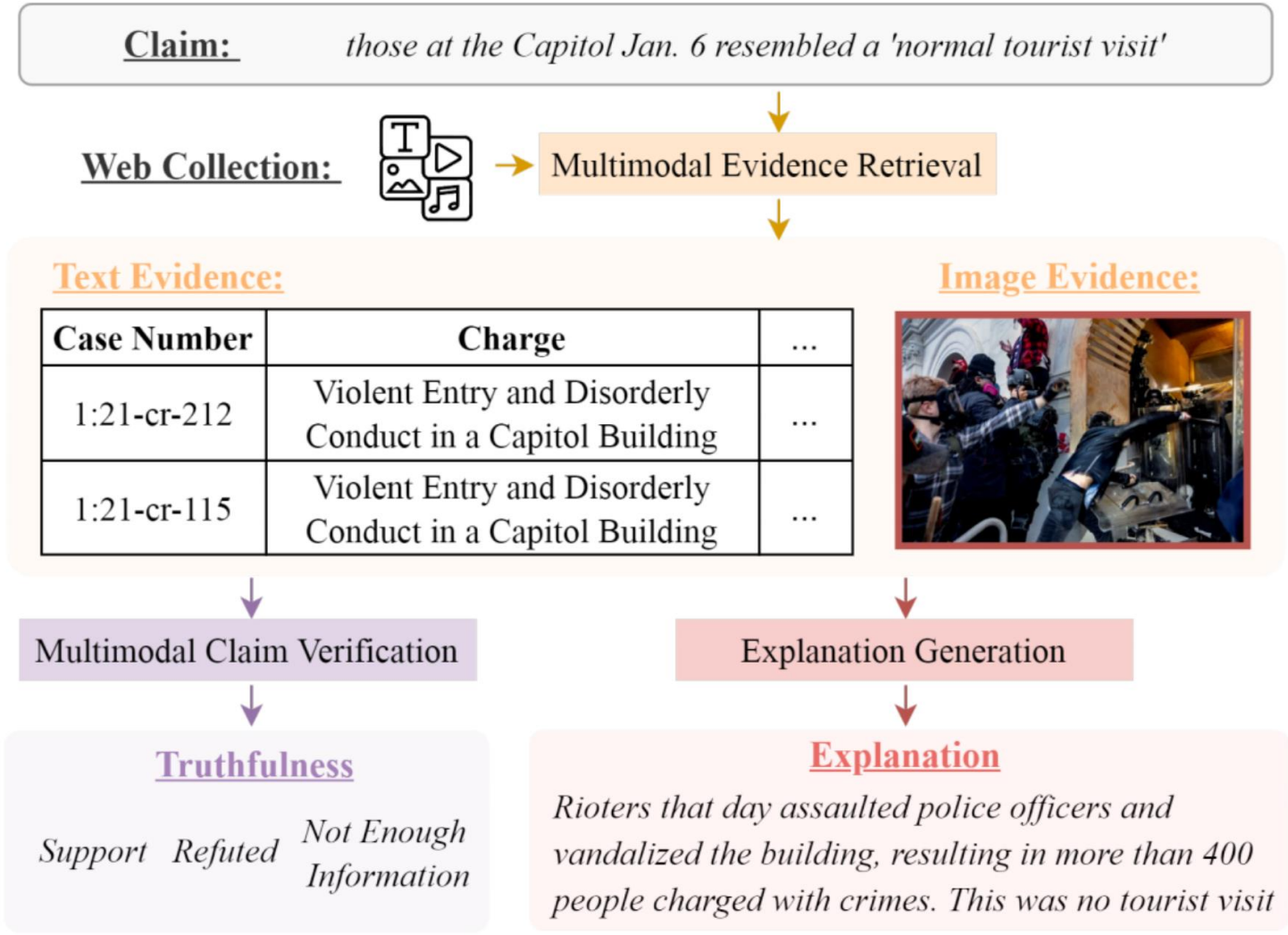


Figure 1: An example of end-to-end multimodal fact-checking and explanation generation.

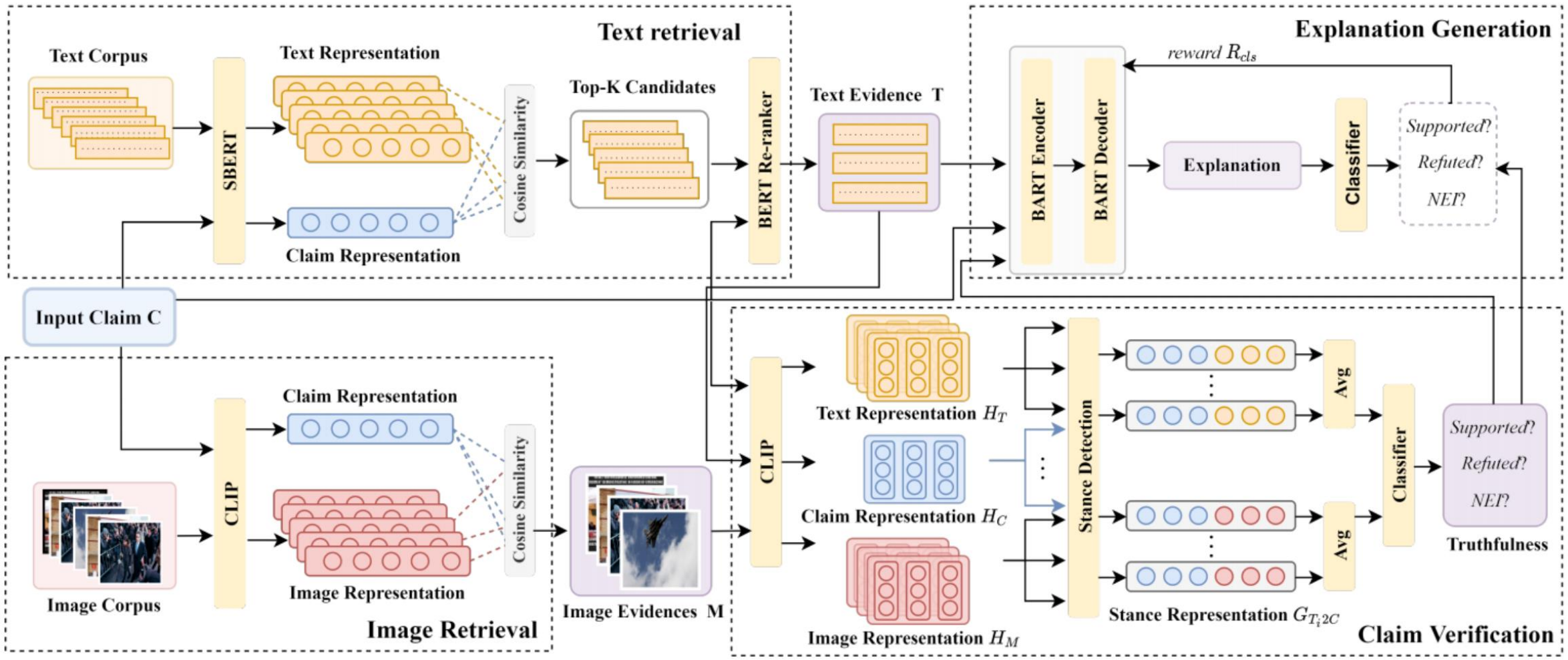
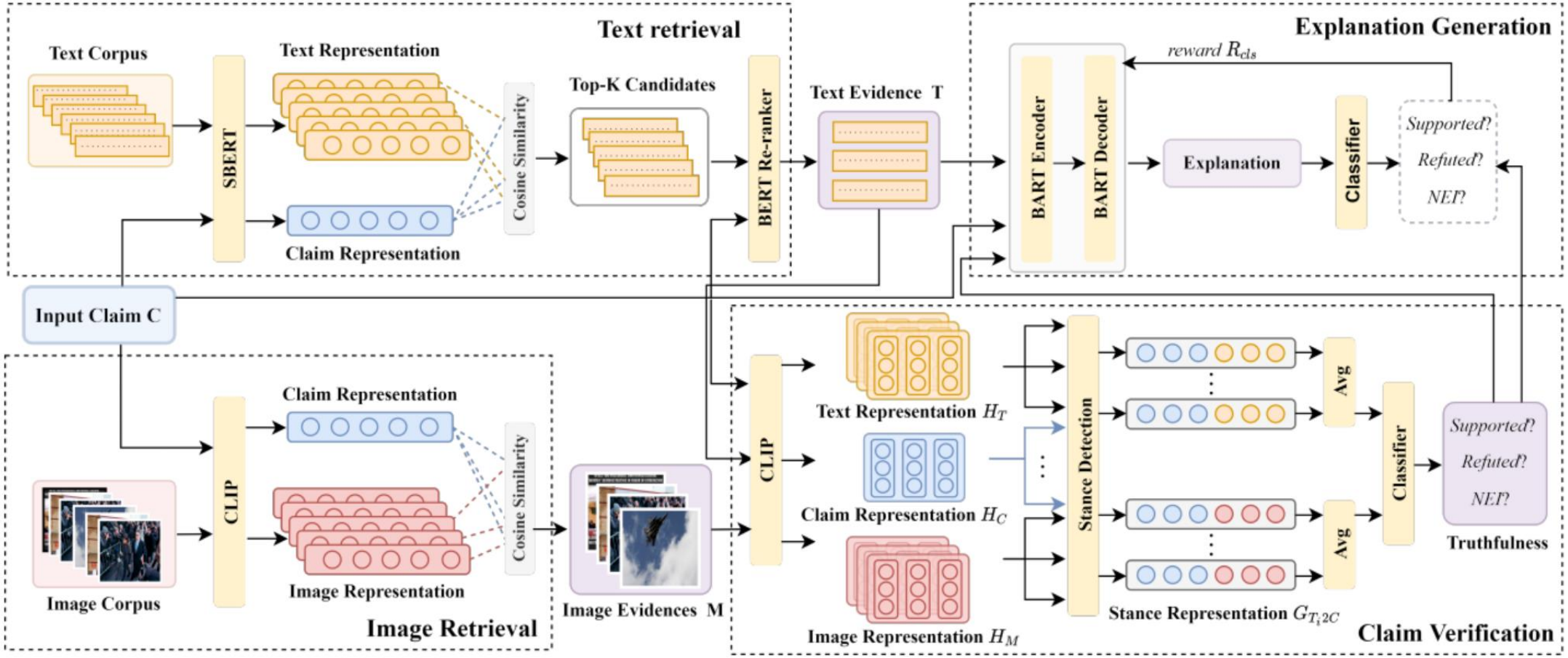


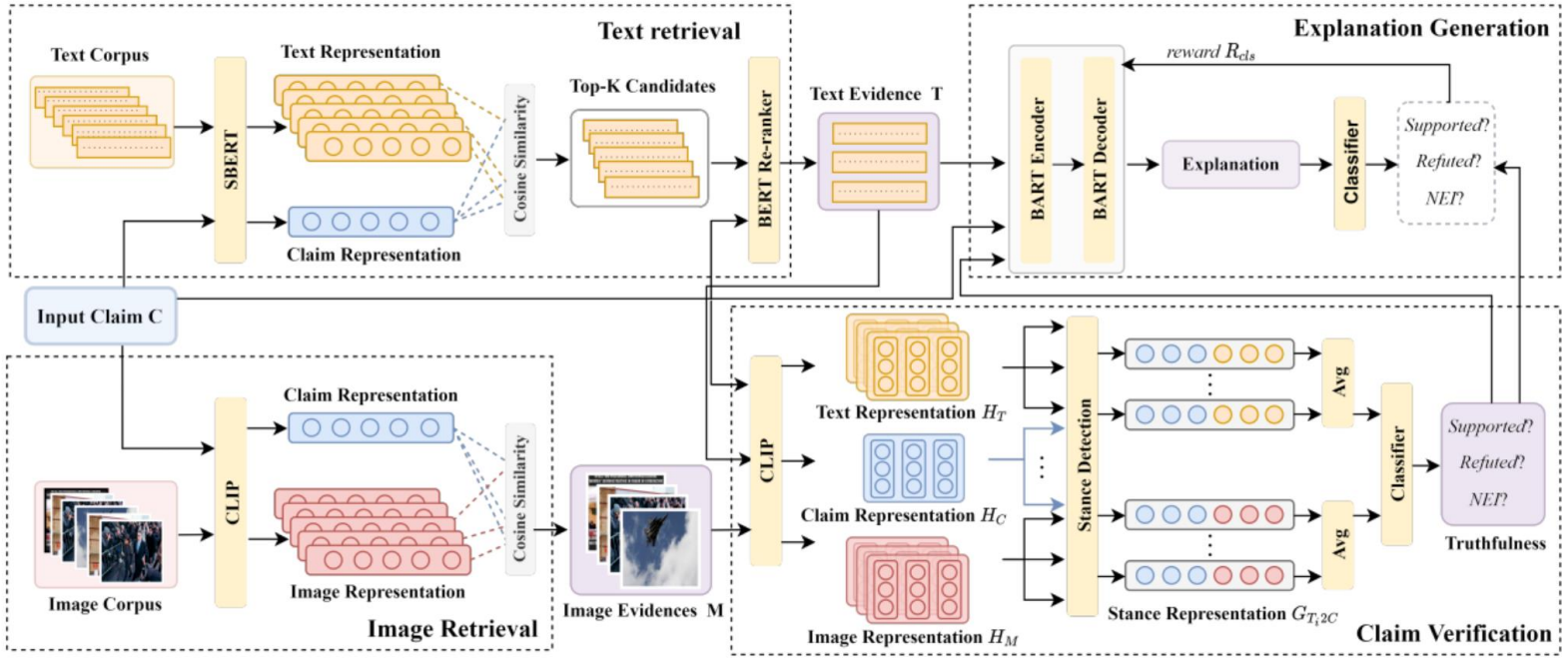
Figure 2: Overview of framework. It consists of a text retrieval module (top left), a image retrieval module(bottom left), a claim verification module(bottom right), and an explanation generation module(top right)

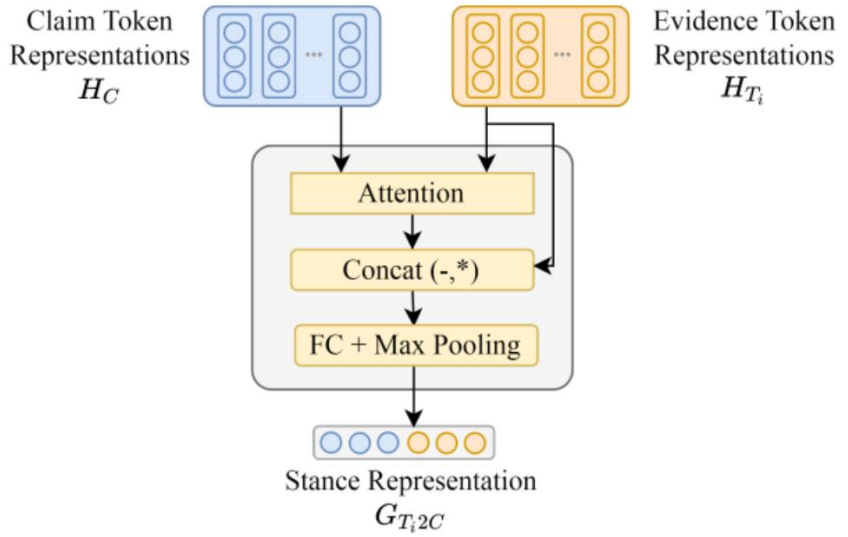
Method



$$C = \{c_0, c_1, ..., c_n\} \quad T_i = \{t_{i0}, t_{i1}, ..., t_{is}\} \quad M_j = \{m_{j0}, m_{j1}, ..., m_{jq}\}$$

$$\boldsymbol{H}_C = \{\boldsymbol{h}_{c_0}, \boldsymbol{h}_{c_1}, ..., \boldsymbol{h}_{c_n}\} \; \boldsymbol{H}_{T_i} = \{\boldsymbol{h}_{t_{i0}}, \boldsymbol{h}_{t_{i1}}, ..., \boldsymbol{h}_{t_{is}}\} \; \boldsymbol{H}_{M_j} = \{\boldsymbol{h}_{m_{j0}}, \boldsymbol{h}_{m_{j1}}, ..., \boldsymbol{h}_{m_{jq}}\}$$





$$\boldsymbol{h}_{\tilde{c}_i} = \text{Softmax}(\boldsymbol{h}_{c_i} \cdot \boldsymbol{H}_{T_i}^{\top}) \cdot \boldsymbol{H}_{T_i}$$

$$\boldsymbol{H}_{T_i2C} = \{\boldsymbol{h}_{\tilde{c}_0}, \boldsymbol{h}_{\tilde{c}_1}, ..., \boldsymbol{h}_{\tilde{c}_n}\} \quad (1)$$

$$\tilde{\boldsymbol{G}}_{T_i2C} = \sigma([\boldsymbol{H}_{T_i2C}\boldsymbol{H}_C : \boldsymbol{H}_{T_i2C} - \boldsymbol{H}_C]\boldsymbol{W}_a + \boldsymbol{b}_a)$$

$$\boldsymbol{G}_{T_i2C} = \text{Max\_Pooling}(\tilde{\boldsymbol{G}}_{T_i2C}) \quad (2)$$

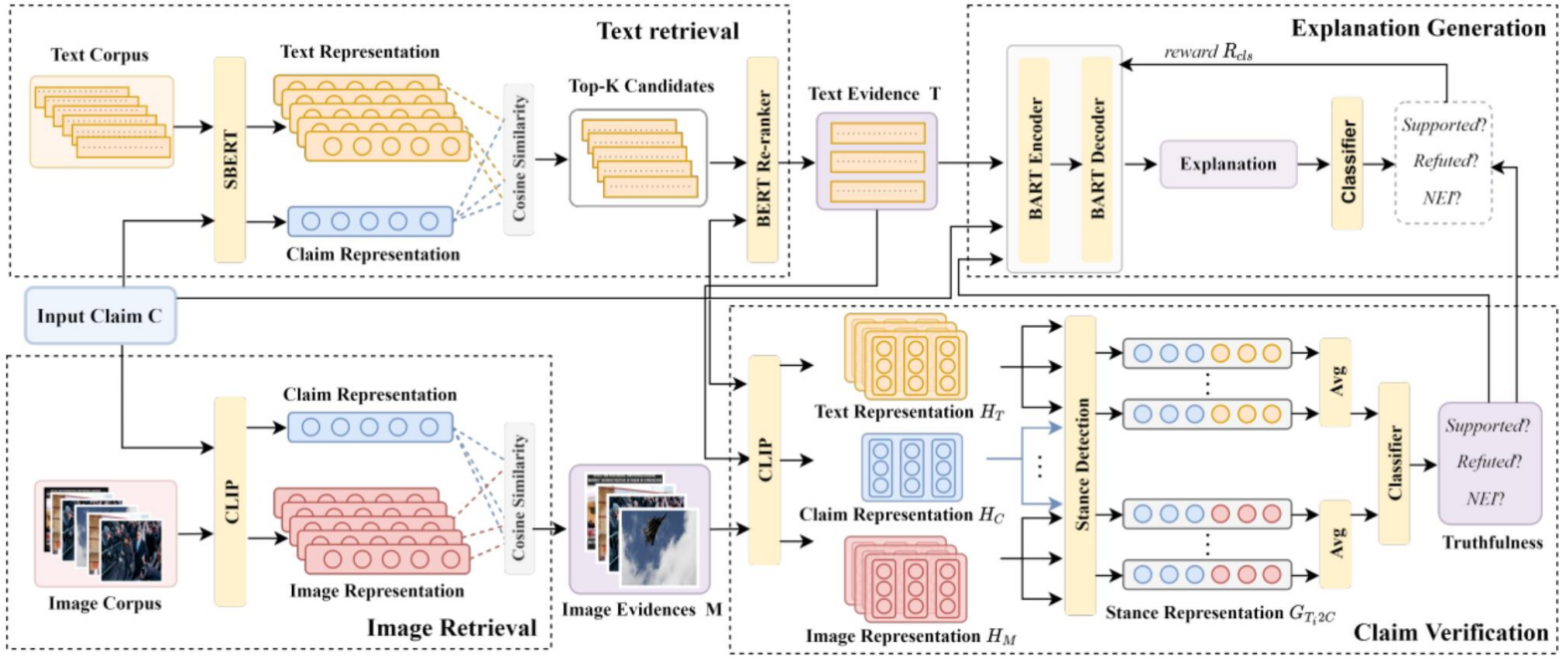
$$\boldsymbol{G}_{T2C} = \text{Mean\_Pooling}(\boldsymbol{G}_{T_i2C})$$

$$\boldsymbol{G}_{M2C} = \text{Mean\_Pooling}(\boldsymbol{G}_{M_j2C})$$

$$\hat{\boldsymbol{y}}_{cls} = \boldsymbol{W}_h^{\top} \cdot [\boldsymbol{G}_{T2C} : \boldsymbol{G}_{M2C}] + \boldsymbol{b}_h$$

$$\mathcal{L}(y_i|C) = -\log\left(\frac{\exp(\hat{\boldsymbol{y}}_{cls,i})}{\sum_{j=0}^{2}\exp(\hat{\boldsymbol{y}}_{cls,j})}\right) \quad (3)$$

Method



Specifically, given an input claim C, its truthfulness label y_C, and text evidences $\{T_1, T_2, ..., T_5\}$, we concatenate them into an overall sequence X

$$S = \{s_1, s_2, ..., s_q\}$$

$$\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_q\}$$

$$\mathcal{L}_g = -\sum_i \log(p(\tilde{s}_i|\tilde{s}_{1:i-1}, X; \phi)) \quad (4)$$

$$\boldsymbol{p}(\tilde{y}|S) = \text{Softmax}_i(\text{classifier}_\theta(S))$$

$$R_{cls} = \boldsymbol{p}(\tilde{y}_C|S) - \sum_{\tilde{y}_j != \tilde{y}_C, \tilde{y}_j \in \{0,1,2\}} \boldsymbol{p}(\tilde{y}_j|S) \quad (5)$$

$$\nabla_\phi \mathcal{J}(\phi) = \mathbb{E}[\lambda \cdot R_{cls} \cdot \nabla_\phi \sum_i \log(\boldsymbol{p}(s_i|s_{1:i-1}, X; \phi))]$$

Experiments

Dataset	Media	re-ranking?	Precision	Recall	F-score
Train	Image	-	58.97	66.14	62.34
Dev	Image	-	60.39	68.97	64.40
Test	Image	-	56.37	64.46	60.14
Train	Text	w/	52.84	37.93	44.16
Dev	Text	w/	52.98	39.61	45.33
Test	Text	w/	53.15	41.22	46.43
Train	Text	w/o	52.46	37.60	43.80
Dev	Text	w/o	52.50	39.39	45.01
Test	Text	w/o	53.12	41.11	46.35

Table 2: Performance of Text and Image Evidence Retrieval on Training, Development, and Test Sets. (%)

Experiments

Setting	F-score
w/o Evidence	33.98
w/ Text Evidence (Gold)	45.18
w/ Image Evidence (Gold)	40.93
w/ Text and Image evidence (Gold)	49.43
w/ Text Evidence (System)	41.03
w/ Image Evidence (System)	38.68
w/ Text and Image evidence (System)	46.78

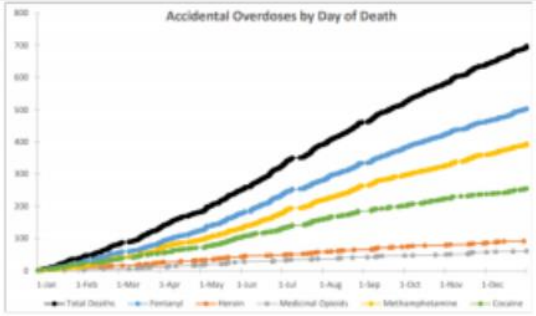
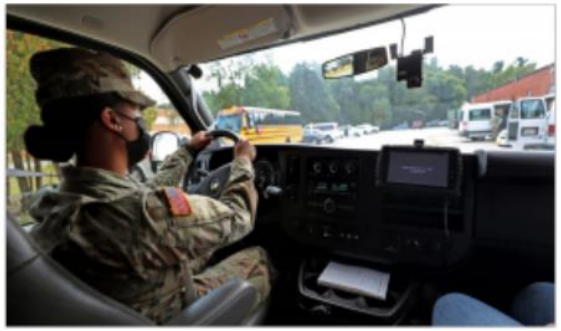
Table 3: Performance of Claim Verification based on Gold and System-retrieved Evidence. (%)

Experiments

Setting	Model	Rouge1	Rouge2	RougeL	BLEU	BERTScore
Gold Evidence w/o Generation	-	36.47	19.04	23.78	16.25	86.60
System Evidence w/o Generation	-	26.36	7.15	15.35	5.11	83.32
Gold Evidence + Gold Truthfulness	BART-large	46.21	26.52	35.59	16.73	86.67
Gold Evidence + System Truthfulness	BART-large	39.93	22.43	27.58	16.70	86.67
System Evidence + Gold Truthfulness	BART-large	28.75	10.73	17.33	7.03	83.31
System Evidence + System Truthfulness	BART-large	28.74	10.72	17.29	7.00	83.31

Table 4: Performance of Explanation Generation. (%)

Experiments

Claim	Text Evidence	Image Evidence	Truthfulness
#1: San Francisco had twice as many drug overdose deaths as COVID deaths last year	That's more than twice San Francisco's 257 deaths due to COVID-19		*Supported*
#2: To address a shortage of school bus drivers in September 2021, Massachusetts Gov. Charlie Baker directed National Guard troops to help transport K-12 students to school	Governor Charlie Baker today will activate the Massachusetts National Guard in response to requests from local communities for assistance with school transportation as the 2021-2022 school year gets underway in the Commonwealth. Beginning with training on Tuesday, 90 Guard members will prepare for service in Chelsea, Lawrence, Lowell, and Lynn		*Supported*
#3: A photograph shows actor Tom Cruise sitting on top of the Burj Khalifa skyscraper without a harness	Special mounts had to be made for the 65 millmeter Imax cameras, special safety had to be put in place, because in a building that's 800 meteres tall [it's 2,723 feet] you couldn't run the risk of anything falling		*Supported*

Thanks