A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection

Zhiwei Yang[1,2,3,5], Jing Ma[2,*], Hechang Chen[3,5,*], Hongzhan Lin[2], Ziyang Luo[2], Yi Chang[3,4,5,*]

[1] College of Computer Science and Technology, Jilin University, Changchun, China
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[3] School of Artificial Intelligence, [4] International Center of Future Science, Jilin University, China
[5] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education

yangzw18@mails.jlu.edu.cn, chenhc@jlu.edu.cn,
{majing, cszyluo, cshzlin}@comp.hkbu.edu.hk, yichang@jlu.edu.cn

COLING2022

code: https://github.com/Nicozwy/CofCED

Introduction

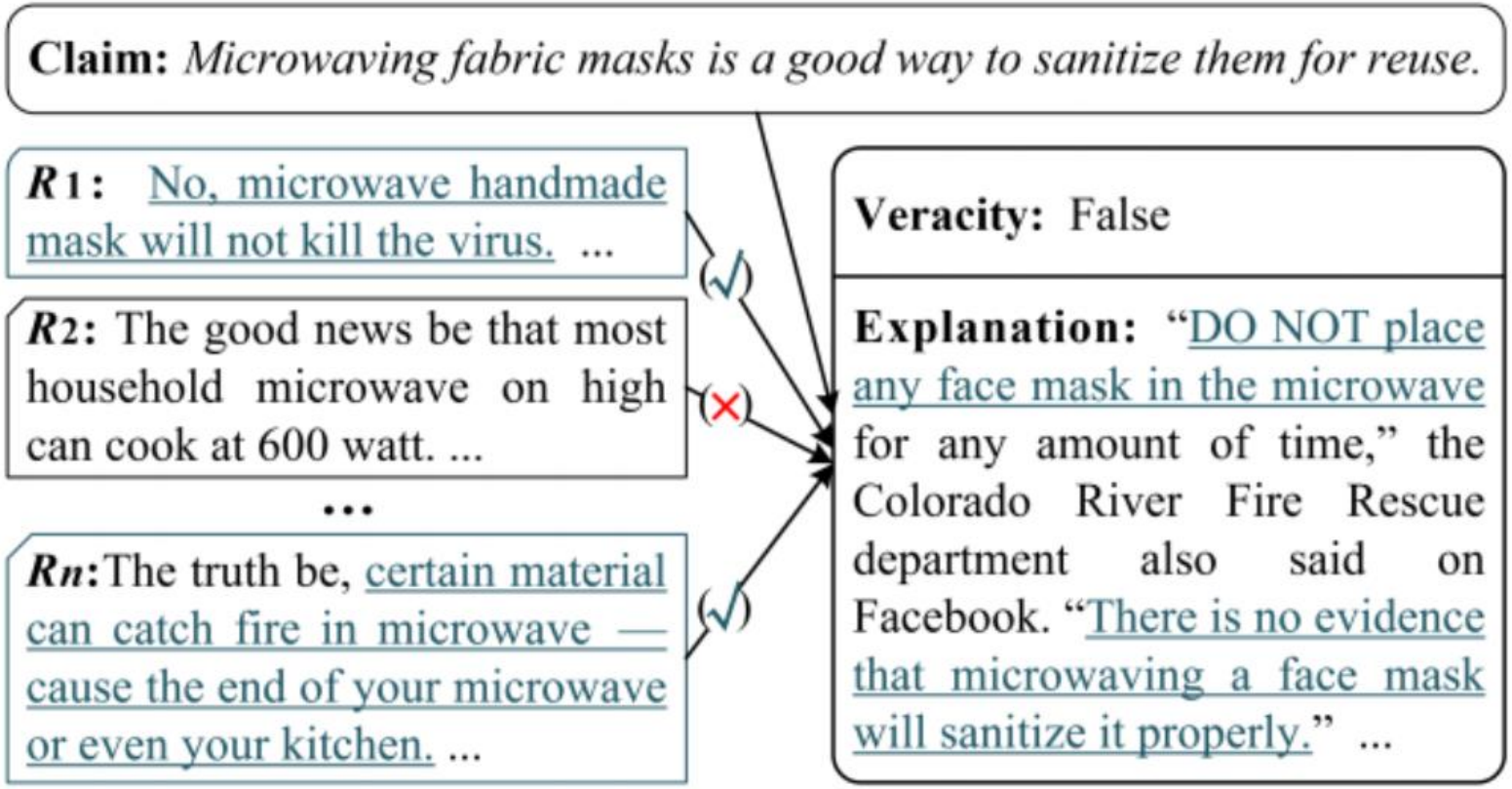


Figure 1: An example for veracity explanation generation. The underlined explanations can be semantically inferred from some relevant sentences in the reports R_1 and R_n. "R" denotes the raw report.

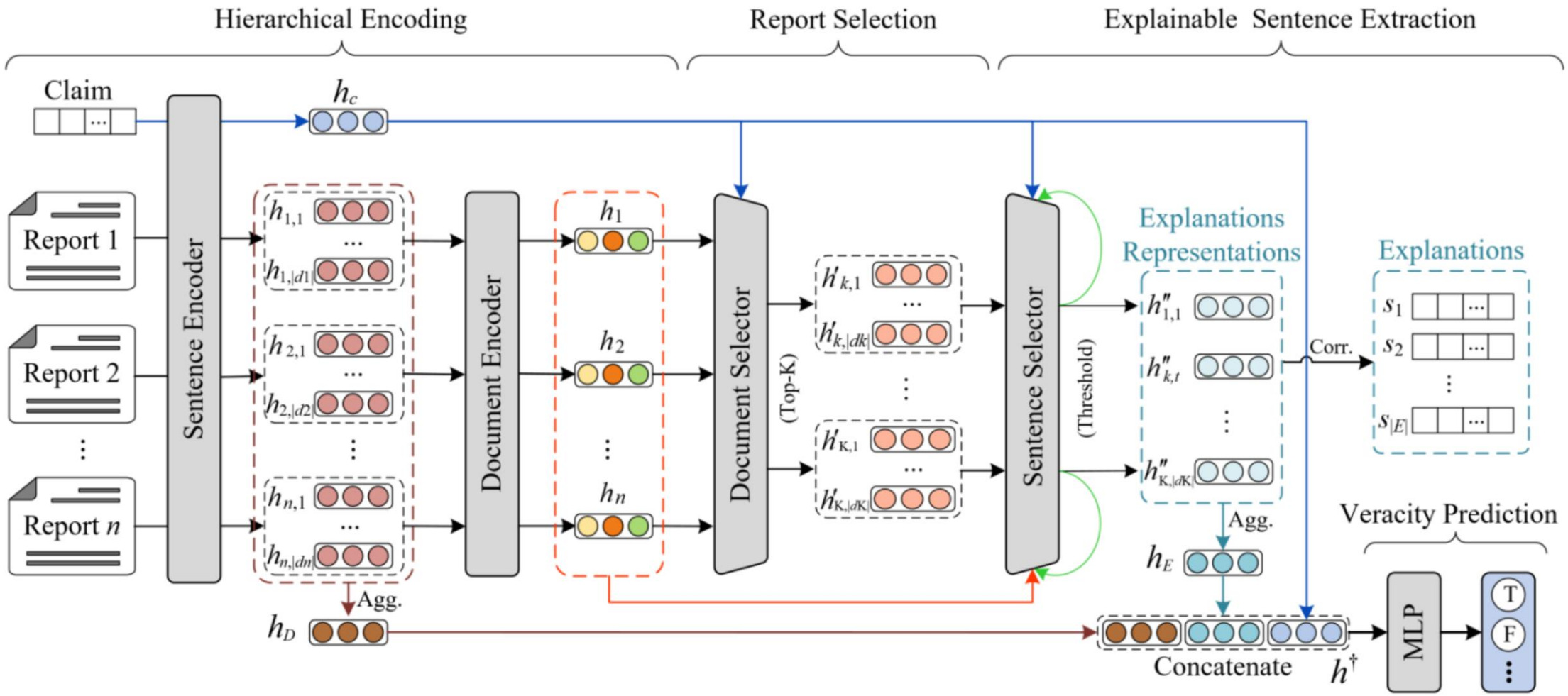
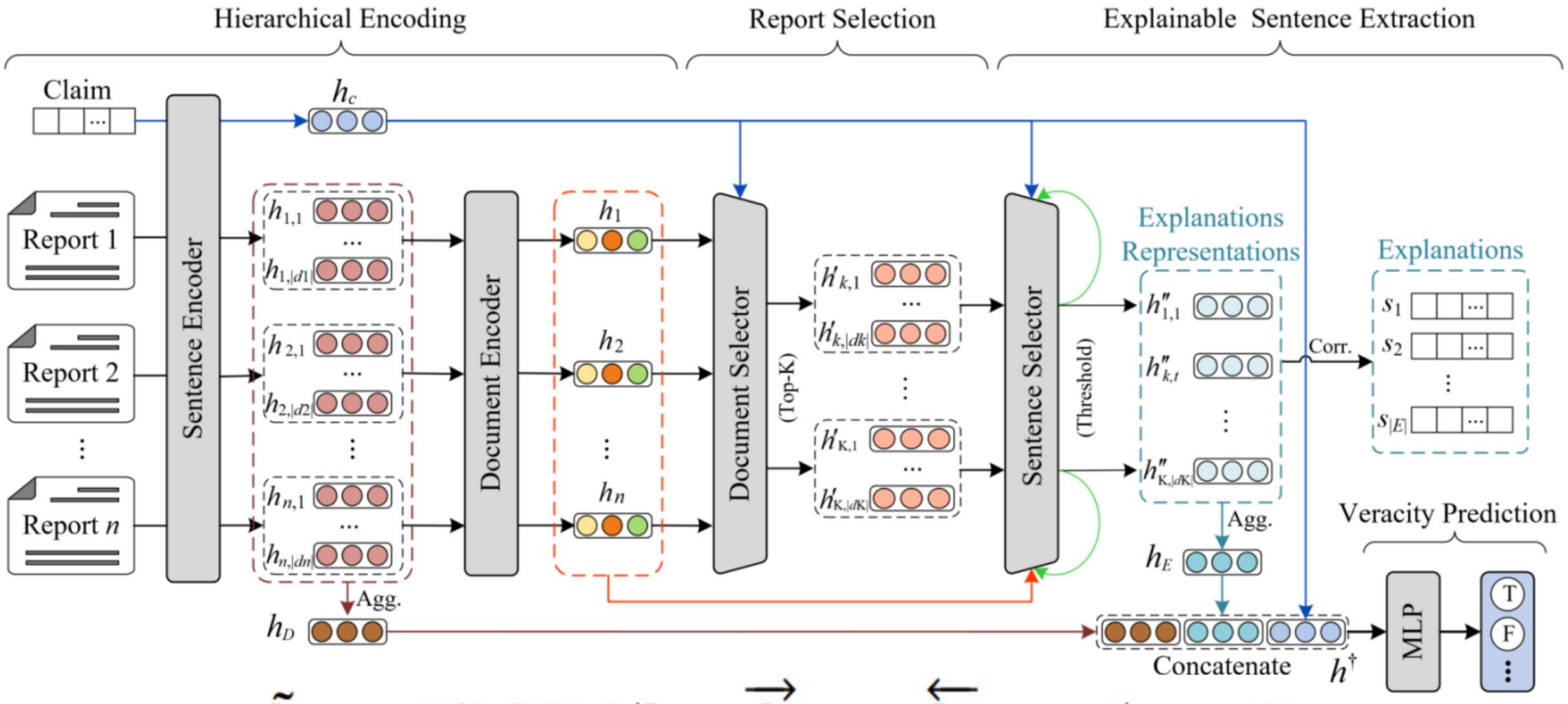


Figure 2: An overview of our proposed CofCED framework. The document selector and the sentence selector are used for selecting check-worthy reports (containing oracles) and oracles, respectively. "Agg." denotes aggregation and " Corr." denotes corresponding. We use different color to highlight different objects. Note that the green line denotes the last output of sentence selection for checking redundancy.

Method



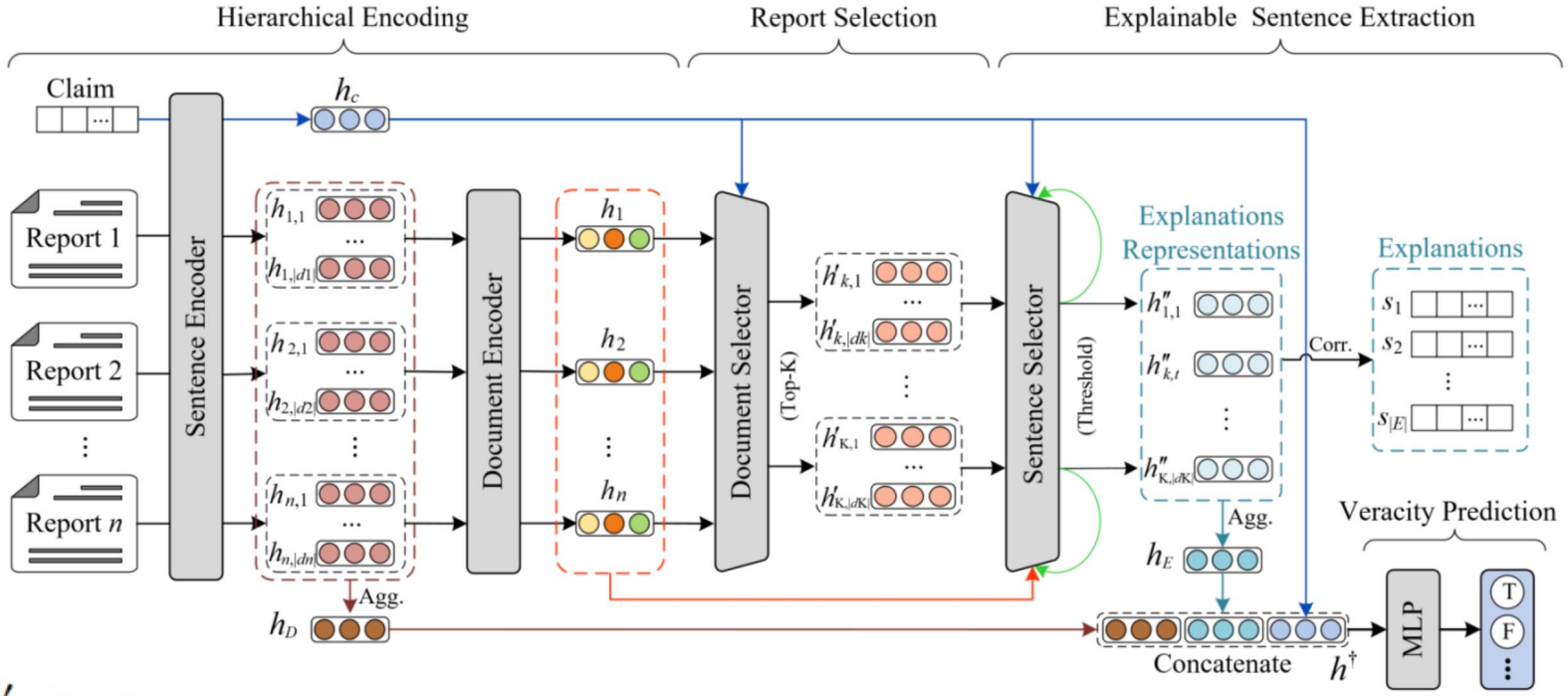
$$\tilde{\mathbf{h}}_{i,j} = \mathrm{BiLSTM}(\mathbf{h}_{i,j}, \overrightarrow{\mathbf{h}}_{i,j-1}, \overleftarrow{\mathbf{h}}_{i,j-1}, \theta) \quad (1)$$

$$\mathbf{h}_i = \mathrm{Max}([\tilde{\mathbf{h}}_{i,1}; \tilde{\mathbf{h}}_{i,2}; \cdots; \tilde{\mathbf{h}}_{i,|d_i|}]) \quad (2)$$

$$\alpha_{c \to \mathcal{D}} = \mathrm{softmax}(\mathbf{H}_{\mathcal{D}} W_\alpha \mathbf{h}_c) \quad (3)$$

$$\mathbf{H}_{\mathcal{D}} = [\mathbf{h}_1; \mathbf{h}_2; \cdots; \mathbf{h}_{|\mathcal{D}|}]$$

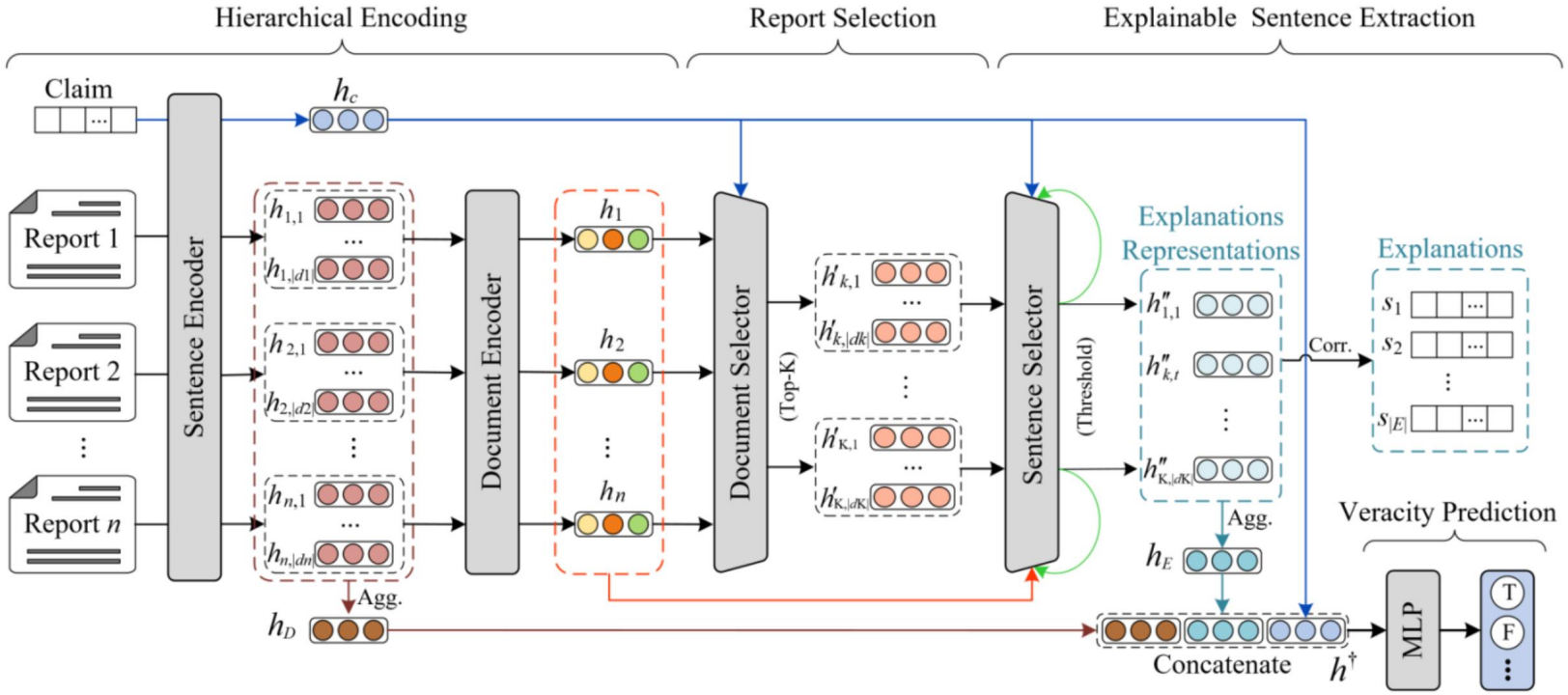
Method



$$\mathrm{P}(y^s_{k,t}=1|\mathbf{h}_c,\mathbf{h}'_{k,t},\mathbf{h}'_k,\mathbf{h}_d)$$
$$=\sigma(\underbrace{\mathbf{h}'_{k,t}W_c\mathbf{h}_c}_{(claim\ relevance)}+\underbrace{\mathbf{h}'_{k,t}W_s}_{(richness)}$$
$$+\underbrace{\mathbf{h}'_{k,t}W_r\mathbf{h}'_k}_{(salience)}-\underbrace{\mathbf{h}'_{k,t}W_d\mathbf{h}_d}_{(non\text{-}redundancy)}) \quad (4)$$

$$\mathbf{h}_d=\tanh(\sum \mathbf{h}'_{k-1,t}\cdot \mathrm{P}(y^s_{k,t}=1)) \quad (5)$$
$$\mathbf{h}_D=\mathrm{Max}([\mathbf{h}_1;\mathbf{h}_2;\cdots;\mathbf{h}_{|\mathcal{D}|}]) \quad (6)$$
$$\mathbf{h}_E=\mathrm{Max}([\mathbf{h}''_1;\mathbf{h}''_2;\ldots;\mathbf{h}''_K]) \quad (7)$$
$$\mathbf{h}^\dagger=[\mathbf{h}_c;\mathbf{h}_D;\mathbf{h}_E] \quad (8)$$
$$\hat{y}=\mathrm{softmax}(\mathrm{MLP}(\mathbf{h}^\dagger)) \quad (9)$$

Method



$$\mathcal{L}_D = -\sum_i y_i^d \log(\hat{y}_i^d) \tag{10}$$

$$\mathcal{L}_S = -\sum_k \sum_t y_{k,t}^s \log(\hat{y}_{k,t}^s) \tag{11}$$

$$\mathcal{L}_C = -y\log(\hat{y}) \tag{12}$$

$$\mathcal{L}_{all} = \beta_D \mathcal{L}_D + \beta_S \mathcal{L}_S + \beta_C \mathcal{L}_C \tag{13}$$

Experiments

Dataset	RAWFC	LIAR-RAW
Claim	2,012	12,590
# pants-fire	-	1,013
# false	646	2,466
# barely-true	-	2,057
# half-true †	671	2,594
# mostly-true	-	2,439
# true	695	2,021
Veracity Label	3	6
Explain sentence		
# min	1	1
# max	110	209
# avg	18.4	4.1
Report per claim		
# min	1	1
# max	30	30
# avg	21.0	12.3
Sentence per report		
# min	1	1
# max	155	59
# avg	7.4	5.5

Table 1: Statistics of datasets. # half-true † is also denoted as # half in RAWFC. The number of oracles in datasets isn't pre-defined.

Experiments

Model	RAWFC			LIAR-RAW		
	P(%)	R(%)	macF1(%)	P(%)	R(%)	macF1(%)
SVM (Pedregosa et al., 2011)	32.33	32.51	31.71	15.78	15.92	15.34
CNN (Wang, 2017)	38.80	38.50	38.59	22.58	22.39	21.36
RNN (Rashkin et al., 2017)	41.35	42.09	40.39	24.36	21.20	20.79
DeClarE (Popat et al., 2018)	43.39	43.52	42.18	22.86	20.55	18.43
dEFEND (Shu et al., 2019)	44.93	43.26	44.07	23.09	18.56	17.51
SentHAN (Ma et al., 2019)	45.66	45.54	44.25	22.64	19.96	18.46
SBERT-FC (Kotonya and Toni, 2020b)	51.06	45.92	45.51	24.09	22.07	22.19
GenFE (Atanasova et al., 2020)	44.29	44.74	44.43	28.01	26.16	26.49
GenFE-MT (Atanasova et al., 2020)	45.64	45.27	45.08	18.55	19.90	15.15
CofCED	**52.99**	**50.99**	**51.07**	**29.48**	**29.55**	**28.93**

Table 2: Experimental results of veracity prediction merely using raw reports ($p < 0.05$ under t-test).

Experiments

Model	RAWFC			LIAR-RAW		
	P(%)	R(%)	macF1(%)	P(%)	R(%)	macF1(%)
CofCED w/o RS&SE	45.01	45.02	44.98	25.69	24.55	24.80
CofCED w/o SE	52.27	46.36	43.80	27.59	23.81	23.74
CofCED w/o RS	49.26	46.92	46.37	27.08	25.32	25.52
CofCED w/o non-redundancy	48.80	46.98	47.48	26.54	27.36	26.65
CofCED w/o salience	43.96	49.24	46.44	26.36	24.88	25.23
CofCED w/o richness	48.08	47.50	47.12	27.06	25.82	26.05
CofCED w/o claim relevance	45.66	45.25	45.28	26.42	24.01	24.88
CofCED	**52.99**	**50.99**	**51.07**	**29.48**	**29.55**	**28.93**

Table 3: Ablation study results of our veracity prediction on test sets; w/o denotes 'without'.

Experiments

Model	RAWFC			LIAR-RAW		
	ROU-1	ROU-2	ROU-L	ROU-1	ROU-2	ROU-L
LEAD-N	19.52	4.54	17.26	9.84	0.40	7.20
Oracle	37.62	13.22	34.67	25.50	9.28	22.61
EXTABS (Kotonya and Toni, 2020b)	-	-	-	18.85	3.61	12.90
dEFEND (Shu et al., 2019)	19.95	5.08	17.21	17.03	3.26	11.42
GenFE-MT (Atanasova et al., 2020)	18.23	7.12	17.32	**23.08**	3.67	12.10
CofCED w/o non-redundancy	27.32	9.06	23.19	17.96	3.54	12.43
CofCED w/o salience	26.67	7.44	21.02	17.27	3.41	11.69
CofCED w/o richness	25.75	8.66	21.87	17.23	3.44	12.10
CofCED w/o claim relevance	25.56	8.07	20.73	17.08	3.31	11.25
CofCED w/o RS	26.64	8.96	22.69	17.51	**3.72**	**13.20**
CofCED	**27.62**	**9.32**	**23.57**	17.14	3.49	12.96

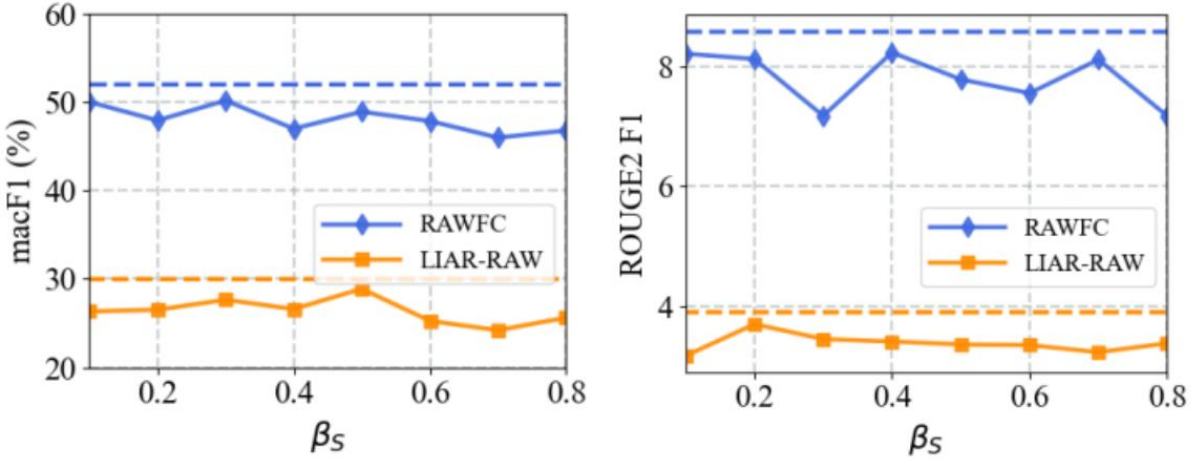
Table 4: ROUGE results of the generated explanation. ROU-N ($N \in \{1, 2, L\}$) denotes the ROUGE-N F1 score that evaluates the token overlap between the explanation and human justifications. RAWFC is not suitable for EXTABS because its gold justification is too long to train an abstractive-summarization model.

Experiments

	Relevance	Richness	Salience	Non-redant	**Overall**	
Claim: *Dr. Tasuku Honjo said that COVID-19 was "man-made" at a lab in Wuhan, China.* [**Prediction:** False] **Explanation:** Honjo did not work at the Wuhan Institute of Virology, he did not say that COVID-19 was "invented" or "man-made," and the Twitter account posting similar claims does not belong to the Nobel Prize winner. In addition, this rumor is all based on the unfounded notion that COVID-19 was created as a bioweapon. (...)						
[1] TOKYO, May 6 (Xinhua) – Japanese Nobel laureate Tasuku Honjo have refuted claim that China manufacture the novel coronavirus, say those rumor be "dangerously distract."	0.9	0.6	0.8	0.9	**0.9**	√
[2] Actually, the professor don't have a Twitter account.	0.7	0.5	0.6	0.9	**0.6**	√
[3] The 2018 Nobel laureate encourage Japanese authority to adopt a more proactive approach.	0.3	0.5	0.4	0.8	**0.3**	×
[4] China will have a big role to play. ...	0.2	0.2	0.1	0.7	**0.2**	×

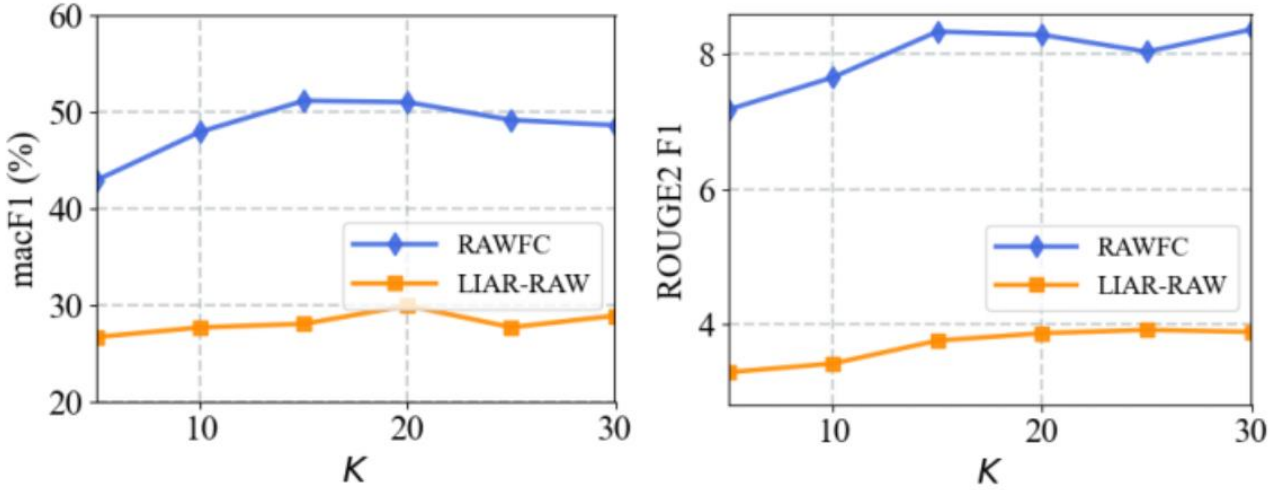
Table 5: Our visualization of explanation extraction from raw reports. Each row is a sentence in raw reports. The score in the columns are normalized from each of the abstract features in Eq. (4), and the last column is the final probability explaining to detection results.

Experiments



(a) Veracity prediction (b) Explanation generation

Figure 3: Results of CofCED under different values of the trade-off parameter β_S and $\beta_C = 1 - \beta_S$. The colored dashed horizontal lines denote the performance of CofCED with our adaptive weighting.



(a) Veracity prediction (b) Explanation generation

Figure 4: Results of CofCED under different values of the maximum number K for report selection.

Thanks