Unified Dual-view Cognitive Model for Interpretable Claim Verification

Lianwei Wu, Yuan Rao, Yuqian Lan, Ling Sun, Zhaoyin Qi
Xi'an Key Lab. of Social Intelligence and Complexity Data Processing,
School of Software Engineering, Xi'an Jiaotong University, China
Shaanxi Joint Key Laboratory for Artifact Intelligence(Sub-Lab of Xi'an Jiaotong University), China
Research Institute of Xi'an Jiaotong University, Shenzhen, China
{stayhungry,yuqian_lan_xjtu,sunling,daqige123}@stu.xjtu.edu.cn
raoyuan@mail.xjtu.edu.cn

Abstract

Recent studies constructing direct interactions between the claim and each single user response to capture evidence have shown remarkable success in interpretable claim verification. Owing to different single responses convey different cognition of individual users, the captured evidence belongs to the perspective of individual cognition. However, individuals' cognition of social things is not always able to truly reflect the objective. There may be one-sided or biased semantics in their opinions on a claim. The captured evidence correspondingly contains some unobjective and biased information. In this paper, we propose a Dual-view model based on the views of Collective and Individual Cognition (CICD) for interpretable claim verification. For collective cognition, we not only capture the word-level semantics based on individual users, but also focus on sentence-level semantics (i.e., the overall responses) among all users to generate global evidence. For individual cognition, we select the top-k articles with high degree of difference and interact with the claim to explore the local key evidence fragments. To weaken the bias of individual cognition-view evidence, we devise an inconsistent loss to suppress the divergence between global and local evidence for strengthening the consistent shared evidence between the both. Experiments on three benchmark datasets confirm the effectiveness of CICD.

1 Introduction

The problem of claim credibility has seriously affected the media ecosystem. Research (Allen et al., 2020) illustrates that the prevalence of 'fake news' has decreased trust in public institutions, and undermined democracy. Meanwhile, 'massive infodemic' during COVID-19 has taken a great toll on health-care systems and lives (Fleming, 2020). Therefore, how to verify the claims spread in networks has become a crucial issue.

Current approaches on claim verification could be divided into two categories: 1) The first category re-

Figure 1: Users' individual cognition of a false claim.

lies on traditional machine learning and deep learning methods to capture semantics (Yang et al., 2019), sentiments (Ajao et al., 2019), writing styles (Przybyla, 2020), and stances (Kumar and Carley, 2019) from claim content, and meta-data features, such as user profiles (Shu et al., 2019; Wu et al., 2020b) for verification. Such approaches could improve verification performance, but they are hard to make reasonable explanations for the verified results, i.e., where false claims go wrong; 2) To tackle this issue, many researchers further focus on interpretable claim verification (the second category) by establishing interactive models between claims and each individual relevant article (or comment) to explore coherent (Ma et al., 2019; Wu et al., 2021), similar (Nie et al., 2019; Wu et al., 2020a), or conflicting (Zhou et al., 2020) semantics as evidence for verifying the false parts of claims.

In interpretable claim verification, the majority of models construct interactions between claims and each single user response (i.e., a comment or a relevant article) to capture evidence, which could effectively learn some of errant aspects of false claims. Due to different single responses reflect the cognition of different individual users, the evidence captured by these models is usually confined to individual cognition. However, individuals' cognition of social things is not always able to truly reflect the objective (Greenwald et al., 1998; Boogert et al., 2018). Owing to individuals are affected by factors such as emotional tendency (Ji et al., 2019), traditional beliefs (Willard and Norenzayan, 2017), and selectively capturing information (Hoffman, 2018), there

are considerable differences in cognition of different individuals, and they are prone to cognitive bias, like primacy effect (Troyer, 2011) and halo effect (Goldstein and Naglieri, 2011), there may be one-sided or biased semantics in their expressed opinions. Thus, the captured evidence also correspondingly contains some unobjective and biased evidence fragments, deteriorating task performance. For instance, as shown in Figure 1, facing a claim to be verified, different individual users (here, users are the normal users on social media, not journalists or professionals) have different reactions. R2 (i.e., response 2 or relevant article 2) and R3 released by users contain unreliable and biased information perceived by their individuals, which may lead to some misleading information being captured as evidence by existing interactive models. Therefore, how to explore users' collective cognition on claims is a major challenge for interpretable claim verification.

To address the deficiencies, we propose a unified **D**ual-view model based on **C**ollective and **I**ndividual **C**ognition (CICD) for interpretable claim verification, which focuses on discovering global evidence and local key evidence, respectively, and then strengthens the consistent shared evidence between the both. Specifically, to explore users' collective cognition to capture global evidence, we design Collective cognition view-based Encoder-Decoder module (CED). CED develops claim-guided encoder that not only learns word-level semantics based on individual user, but also captures sentence-level semantics (i.e., the overall opinions) among all users. Here, a relevant article (a response) released by an individual user is usually a sentence sequence, so all sentence-level semantics convey the overall opinions of all users. Then, CED develops hierarchical attention decoder to generate global evidence by adjusting weights of word-level and sentence-level semantics. To further acquire the local key evidence based on individual cognition, we develop Individual cognition view-based Selected Interaction module (ISI) to screen representative top-k articles with high difference and interact with the claim to gain local key evidence fragments. To weaken the bias of individual cognition view and strengthen the consistent shared evidence between global and local evidence, we project inconsistent loss to suppress the divergence. Experimental results not only reveal the effectiveness of CICD but also provide its interpretability. Our contributions are summarized:

- A novel framework integrating interdisciplinary knowledge on interpretable claim verification is explored, which discovers global and local evidence from the perspectives of collective and individual cognition to interpret verified results.

- Proposed CED captures word-level (individual) and sentence-level (holistic) opinions, and reasonably adjusts the proportion between them, which generates global evidence of the view of all users.

- Experiments on three competitive datasets demonstrate that CICD achieves better performance than other strong baselines.

2   Related Work

Automatic verification approaches rely on neural networks to extract content-based features, like semantics (Popat et al., 2018; Wu et al., 2019), sentiments (Nguyen et al., 2020), writing styles (Przybyla, 2020), etc., and metadata-based features, like user profiles-based (Kumar and Carley, 2019), comment-based (Bovet and Makse, 2019), etc., for verification. These methods could improve the accuracy of claim verification, but they are lack of interpretability for the verified results. To tackle this, interpretable claim verification has received great attention. Its basic principle is to obtain queried, corrected, and rumor-refuted semantics from the articles (or comments) related to claims to interpret the credibility of claims. At present, the methods for this task generally focus on direct interactions between claims and relevant articles to identify their matching degree (Nie et al., 2019), consistency (Ma et al., 2019), implication (Liu et al., 2019), conflict (Wu et al., 2020c), etc., to learn practical evidence. For instances, HAN (Ma et al., 2019) and EHIAN (Wu et al., 2020c) learned implication relationships between claims and relevant articles to capture semantic conflicts as evidence, which reflected a certain interpretability. However, since all relevant articles are involved, the captured conflicts may be affected by some low-quality articles with noisy semantics, easily resulting in the invalidation of the evidence. In our model, we design ISI module to screen all relevant articles to capture the valuable representative articles with differential semantics, so as to learn local key evidence fragments. In addition, some methods, such as GEAR (Zhou et al., 2019) and KGAT (Liu et al., 2020), relied on graph-based networks to conduct semantic aggregation and reasoning on relevant articles, so as to capture global evidence. Nevertheless, these models treat an entire article (at the sentence level) as a node and ignore the importance of word-level semantics in each article.

To overcome these defects, our model constructs a hierarchical attention decoder to fuse sentence-level and word-level semantics for finely-grained generating global evidence.

3 The Proposed Approach

In this section, we introduce the details of CICD as illustrated in Figure 2.

Inputs and Outputs For cognitive input representations, the inputs of CED are a claim sequence and the concatenation of its all relevant articles with the number of N, while the inputs of ISI are a claim sequence and each relevant article. Given any a sequence of length l words $\mathbf{X} = \{x_1, x_2, ..., x_l\}$, where each word $x_i \in \mathbb{R}^d$ is a d-dimensional vector obtained by pre-trained BERT model (Devlin et al., 2019). Particularly, the length of each sequence in relevant articles is l and that of the claim sequence is p. Thus, we obtain the representations of the i-th relevant article and the claim as $\mathbf{X}_i^r \in \mathbb{R}^{l \times d}$, $\mathbf{X}^c \in \mathbb{R}^{p \times d}$, respectively. For the outputs of the model, the outputs of CED are the generated global evidence sequence of length o words $\mathbf{G} = \{g_1, g_2, ..., g_o\}$, where g_t is the representation of the t-th generated word and o is the length of $\mathbf{G}$. The outputs of ISI are the integrated vector of top-k local key evidence fragments $\mathbf{I} = [\mathbf{I}_1; \mathbf{I}_2, ; ...; \mathbf{I}_k]$, where ; is the concatenation operation.

3.1 Collective Cognition View-based Encoder-Decoder (CED)

To explore users' collective cognition on claims, we first rely on claim-guided encoder to capture word-level and sentence-level semantics from all relevant articles, and then adjust the proportion between the both by hierarchical attention decoder to generate global evidence.

3.1.1 Claim-guided Encoder

The claim-guided encoder module involves a sequence encoding layer and a matching layer.

Sequence Encoding Layer We rely on BiL-STMs to encode all relevant articles and the claim for their contextual representations. We utilize the produced hidden states $\mathbf{H}^r = \{h_1^r, h_2^r, ..., h_{l_{all}}^r\}$ (where l_{all} means the total length of all articles) and $\mathbf{H}^c = \{h_1^c, h_2^c, ..., h_p^c\}$ to denote the contextual representations of relevant articles and the claim, respectively, where h_i (i.e., h_i^r or h_i^c) is defined as follows:

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \quad (1)$$

where $\overrightarrow{h_i} \in \mathbb{R}^{d_h}$ and $\overleftarrow{h_i} \in \mathbb{R}^{d_h}$ are the i-th hidden state of the forward and backward LSTMs for the word x_i respectively. ; is concatenation operation.

Attention-based Matching Layer is engaged to aggregate the relevant information from the claim for each word within the context of relevant articles. The aggregation operation $a_i = \text{attn}(h_i^r, \mathbf{H}^c)$ is as follows:

$$a_i = \sum_{j=1}^{k_c} \alpha_{i,j} h_j^c \quad (2)$$

$$\alpha_{i,j} = \exp(s_{i,j}) / \sum_{k=1}^{p} \exp(s_{i,k}) \quad (3)$$

$$s_{i,j} = (h_i^r)^\top \mathbf{W}_1 h_j^c \quad (4)$$

where a_i is the aggregated vector for the i-th word of the articles. $\alpha_{i,j}$ is the normalized attention score between h_i^r and h_j^c.

Here, the purpose of adopting claim to guide the encoding of relevant articles includes two perspectives: 1) Strengthening the focus of consistent semantics associated with the claim in relevant articles, i.e., exploring how relevant articles evaluate the claim; and 2) Making the encoding semantics purer. We observe that there are some advertisements or useless information in relevant articles. This way is able to effectively filter the noise irrelevant to the claim from relevant articles, and consolidates the generation of relevant semantics in the decoder module.

Furthermore, we output the hidden state corresponding to the last word encoded by each relevant article to form consistent sentence-level representations, where h_i^s represents sentence-level representations of the i-th relevant article. Particularly, we apply word-level representations $\mathbf{H}^r = \{h_1^r, h_2^r, ..., h_{l_{all}}^r\}$ (which can also be represented in the form of different relevant articles, i.e., $\mathbf{H}^r = \{h_{1,1}^r, h_{1,2}^r, ..., h_{N,l}^r\}$, where $l_{all} = N \times l$) and sentence-level representations $\mathbf{H}^{rs} = \{h_1^s, h_2^s, ..., h_N^s\}$ as memory bank for decoder generation.

3.1.2 Hierarchical Attention Decoder

To capture the collective cognition-view evidence from relevant articles, we devise hierarchical attention decoder to consider the consistent semantics with different granularity of relevant articles to generate global evidence. Specifically, we employ unidirectional LSTM as the decoder, and at each decoding time-step, we calculate in parallel both sentence-level attention weight β and word-level α by:

$$\beta_i = (h_i^s)^\top \mathbf{W}_2 h_t^d \qquad \alpha_{i,j} = (h_{i,j}^r)^\top \mathbf{W}_3 h_t^d \quad (5)$$

$$\gamma_{i,j} = \frac{\alpha_{i,j} \beta_i}{\sum_{i,j} \alpha_{i,j} \beta_i} \quad (6)$$
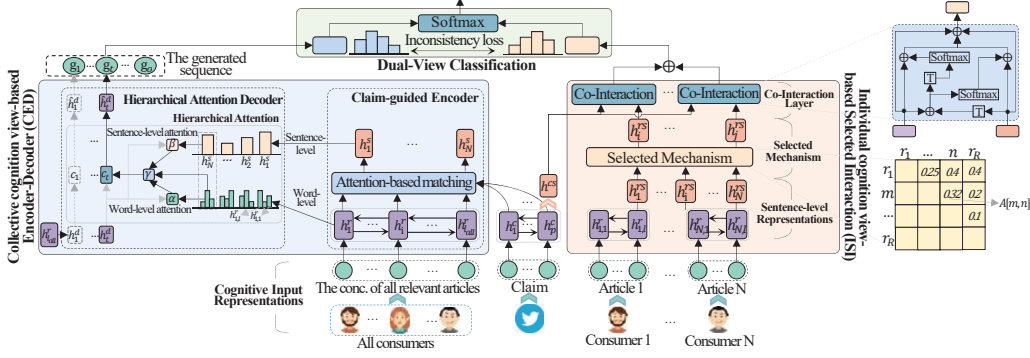
61

Figure 2: The general architecture of CICD. The model consists of: 1) CED for generating global evidence; 2) ISI for exploring local key evidence fragments; and 3) Dual-view classification module leveraging inconsistency loss to promote the learning of shared available evidence between global and local evidence.

where h_t^d is the hidden state of the decoder at the t-th time-step, $\mathbf{W}_2$ and $\mathbf{W}_3$ are trainable parameters. The word-level attention ascertains how to distribute the attention over words in each sentence (each article), which could learn salient evidence segments in each article, while the sentence-level attention determines how much each article should contribute to the generation at current time-step, which could capture potential global semantics in all articles.

Then the context vector c_t is derived as a combination of all word-level representations reweighted by combined attention γ:

$$\mathrm{c}_t = \sum_{i,j} \gamma_{i,j} \mathrm{h}_i^r \qquad (7)$$

And the attentional vector is calculated as:

$$\hat{\mathrm{h}}_t^d = \tanh(\mathbf{W}_4[\mathrm{h}_t^d; \mathrm{c}_t]) \qquad (8)$$

Finally, the predicted probability distribution over the vocabulary V at the current step is:

$$P_V = \mathrm{softmax}(\mathbf{W}_V \hat{\mathrm{h}}_t^d + \mathbf{b}_V) \qquad (9)$$

where $\mathbf{W}_4$, $\mathbf{W}_V$, and $\mathbf{b}_V$ are trainable parameters.

We adopt $\mathbf{G} = \{g_1, g_2, ..., g_o\}$ to denote the generated sequence rich in global evidence.

3.2 Individual Cognition View-based Selected Interaction (ISI)

To capture evidence fragments from individual cognition view, we design ISI module with the following layers: 1) Sentence-level representation for capturing high-level representations of relevant articles; 2) Selected mechanism for screening the representative top-k relevant articles with degree of difference; and 3) Co-interaction layer for making the claim and the selected articles interact with each other to explore local key evidence fragments.

3.2.1 Sentence-level Representation

We exploit BiLSTM to encode each relevant article and capture the output of the last hidden state as the sentence-level representation, where the encoding process is similar to sequence encoding layer in Section 3.2.1, where the sentence-level representation of the i-th article is h_i^{rs}.

3.2.2 Selected Mechanism

To capture representative top-k articles, we develop selected mechanism to calculate the difference between each articles and other articles in an automated manner. To do this, selected mechanism learns and optimizes an inter-sentential attention matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. The entry (m, n) of $\mathbf{A}$ holds the difference between article m and article n ($1 \leq m, n \leq N$ and $m \neq n$) and is computed as:

$$\mathrm{u}_m = \varphi(\mathbf{W}_m \mathrm{h}_m^{rs} + \mathbf{b}_m) \quad \mathrm{u}_n = \varphi(\mathbf{W}_n \mathrm{h}_n^{rs} + \mathbf{b}_n) \quad (10)$$

$$\mathbf{A}[m, n] = \frac{\exp(\mathrm{u}_m \odot \mathrm{u}_n)}{\sum_{i=1}^N \exp(\mathrm{u}_i \odot \mathrm{u}_n)} \qquad (11)$$

where φ is a activation function, $\mathbf{W}_m$ and $\mathbf{W}_n$ are weight matrix, $\mathbf{b}_m$ and $\mathbf{b}_n$ are biases, and $\odot$ denotes dot product operator. The larger the entry $\mathbf{A}[m, n]$ is, the higher the similar between article m and article n is. Thus, the smaller $\mathbf{A}[m, n]$ corresponds to article m and n contain more differential semantics, and finally we screen top-k relevant articles with high difference for further downstream interaction.

3.2.3 Co-Interaction Layer

This co-interaction layer aims to explore local key evidence fragments. Specifically, the layer enables the claim to focus on the i-th article to discover the specific evidence fragment, while the i-th article pays close attention to the claim to explore the possible false part of

the claim. Finally, we combine the two interactions to constitute the individual key local evidence fragments.

$$\mathbf{H}_i^{rin} = \mathrm{h}_i^{rs} + \mathrm{softmax}(\mathrm{h}_i^{rs}((\mathbf{H}^{cs})^\top))\mathbf{H}^{cs} \quad (12)$$

$$\mathbf{H}^{cin} = \mathbf{H}^{cs} + \mathrm{softmax}(\mathbf{H}^{cs}((\mathrm{h}_i^{rs})^\top))\mathrm{h}_i^{rs} \quad (13)$$

$$\mathbf{I}_i = [\mathbf{H}_i^{rin}; \mathbf{H}^{cin}] \quad (14)$$

where $\mathbf{H}_i^{rin}$ is the evidence fragment of the i-th article, $\mathbf{H}^{cin}$ is the false part of the claim, and $\mathbf{H}^{cs}$ is the outputs of the last time step of $\mathbf{H}^c$.

For all top-k articles, we integrate all local evidence fragments by concatenation operation.

$$\mathbf{I} = [\mathbf{I}_1; \mathbf{I}_2; ...; \mathbf{I}_k] \quad (15)$$

3.3 Dual-View Classification

To alleviate the bias of individual cognition-view evidence fragments and strengthen the consistent shared evidence between global and local evidence, we introduce an inconsistency loss to penalize the disagreement between the both evidence. We define the inconsistency loss function as the Kulllback-Leibler (KL) divergence between $\mathbf{G}$ and $\mathbf{I}$.

$$\mathrm{Loss}_{in} = \mathrm{D}_{\mathrm{KL}}(\mathbf{G}||\mathbf{I}) = \sum_{k=1}^{K} \mathbf{G}_k' \log \frac{\mathbf{G}_k'}{\mathbf{I}_k'} \quad (16)$$

where $\mathbf{G}_k'$ is the k-th element of the concatenation of the words in $\mathbf{G}$, and $\mathbf{I}_k'$ is the k-th element of $\mathbf{I}$.

Furthermore, we fuse the two types of penalized evidence, and adopt softmax function to emit the probability distribution for training, where a loss forces the model to minimize the cross-entropy error for a training sample with ground-truth label y:

$$\mathrm{Loss} = -\sum y \log p \quad (17)$$

$$p = \mathrm{softmax}(\mathbf{W}_p[\mathbf{G}; \mathbf{I}] + \mathbf{b}_p) \quad (18)$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the learnable parameters.

To ensure the effective synergy of the two cognition views, we put together all loss mentioned above for joint training.

$$\mathcal{L} = \mathrm{Loss} + \alpha \mathrm{Loss}_{in} \quad (19)$$

where α is the hyper-parameter.

4 Experiments

4.1 Datasets and Evaluation Metrics

For evaluation, we utilize three publicly available datasets, i.e., Snopes, PolitiFact (both released by (Popat et al., 2018)), and FEVER (Thorne et al., 2018). The first two datasets contain 4,341 and 3,568 news claims, associating with 29,242 and 29,556 relevant articles (these articles can be regarded as

responses of different individual users to claims) collected from various web sources respectively. FEVER consists of 185,445 claims accompanied by manual annotation Wikipedia articles. For labels, each claim in Snopes is labeled as $true$ and $false$, while PolitiFact divides claims into six kinds of credibility labels: true, mostly true, half true, mostly false, false, and pants on fire. To distinguish the veracity more practically, like Ma et al. (2019), we merge mostly true, half true and mostly false into mixed, and treat false and pants on fire as false. Then, the labels of PolitiFact are classified as $true$, $mixed$, and $false$. On FEVER, each claim is partitioned as $supported$, $refuted$, or NEI (not enough information). For evaluation metrics, on Snopes and PolitiFact, we exploit micro-/macro-averaged F1(micF1/macF1), class-specific precision (Prec.), recall (Rec.) and F1-score (F1) as evaluation metrics. We hold out 10% of the claims for tuning the hyper parameters, and conduct 5-fold cross-validation on the rest of the claims. On FEVER, we leverage accuracy (Acc.), and F1-score (F1) as evaluation metrics, and follow Thorne et al. (2018) to partition the annotated claims into training, development (Dev.), and testing (Test.) sets.

4.2 Settings

For parameter configurations, we adjust them according to the performance of development sets, we set the word embedding size d to 768. The dimensionality of LSTM hidden states d_h is 120. The length l of each relevant article is 100 and that of the claim p is assigned as 20. Due to no parameters depend on the number of articles N, instead of intercepting a fixed number, we set N to vary with claims. Initial learning rate is set to 2e-3. The loss weight coefficient α is trained to 0.2. The dropout rate is 0.4, and we set the mini-batch size of the three datasets as 32, 32, and 64, respectively. Additionally, an Adam (Kingma and Ba, 2015) optimizer with $\beta 1$ as 0.9 and $\beta 2$ as 0.999 is used to optimize all trainable parameters.

4.3 Experiments on Snopes and PolitiFact

4.3.1 Performance Comparison

We compare CICD and several competitive baselines: 1) **DeClarE** (Popat et al., 2018) models joint interactions between claims and articles and aggregates word-level credibility signals from external articles for evidence-aware assessment; 2) **BERT** (Devlin et al., 2019), we employ pre-trained BERT classifier to verify claims; 3) **HAN** (Ma et al., 2019), a hierarchical attention network, constructs the interactions between claims and relevant articles for capturing sentence-

Methods	Snopes								PolitiFact				
			True			False					True	False	Mixed
	micF1	macF1	Prec.	Rec.	F1	Prec.	Rec.	F1	micF1	macF1	F1	F1	F1
DeClarE	0.762	0.695	0.559	0.556	0.553	0.839	0.837	0.837	0.475	0.443	0.447	0.576	0.307
BERT	0.781	0.706	0.587	0.601	0.594	0.852	0.854	0.853	0.493	0.462	0.478	0.596	0.320
HAN	0.807	0.759	**0.637**	0.665	0.651	0.874	0.860	0.867	0.523	0.487	0.495	0.627	0.340
HAN-ba	0.771	0.738	0.556	0.765	0.644	**0.899**	0.774	0.832	0.520	0.471	0.475	0.629	0.308
EHIAN	0.831	0.784	0.614	0.790	0.691	0.893	0.896	0.894	0.554	0.509	0.513	0.651	0.362
CICD (Ours)	**0.846**	**0.795**	0.629	**0.796**	**0.703**	0.897	**0.904**	**0.900**	**0.572**	**0.525**	**0.529**	**0.665**	**0.373**

Table 1: Comparison of our model with baselines on Snopes and PolitiFact.

level evidence by considering their topical coherence and semantic inference strength; 4) **HAN-ba** (Ma et al., 2019) is a variant of HAN, where the gated attention is replaced to biaffine attention for acquiring evidence; and 5) **EHIAN** (Wu et al., 2020c) is an evidence-aware hierarchical interactive attention network, which focuses on the direct interaction between claim and relevant articles to explore key evidence fragments. As shown in Table 1, we observe that:

- BERT achieves at least 6.5% improvement on micF1 than DeClarE, which illustrates pre-trained model can learn rich semantic context features to improve performance, which is also the reason that we adopt BERT to train word embeddings. HAN consistently outperforms BERT, which indicates HAN capturing the coherence between relevant articles could help improve the task performance.

- In interpretable methods, CICD outperforms De-ClarE, which is because our model not only focuses on word-level semantics like DeClarE, but also grasps the holistic sentence-level features. More-over, owing to HAN and HAN-ba drive all relevant articles to participate in the interaction, prompting them to gain a small boost in precision on Snopes, but this way may introduce noise from nonsignifi-cant articles. CICD effectively avoids this problem by selecting vital articles for interaction, which obtains significant improvements in other metrics compared with HAN and HAN-ba. Furthermore, CICD consistently outperforms EHIAN on Snopes and PolitiFact. The superiority is clear: CICD not only values individual cognition view to capture key evidence fragments, but also generates collec-tive cognition-view evidence for claim verification.

4.3.2 Ablation Study

In order to evaluate the impact of each component of CICD, we ablate CICD into the following simplified models: 1) **-matching U** represents the attention-based matching layer of CED is removed; 2) **-CED**

Methods	Snopes		PolitiFact	
	micF1	macF1	micF1	macF1
-matching U	0.802	0.753	0.529	0.486
-CED	0.791	0.748	0.526	0.476
-selected I	0.810	0.763	0.541	0.490
-interaction I	0.822	0.770	0.557	0.497
-ISI	0.803	0.751	0.530	0.483
-inconsistency loss	0.831	0.782	0.556	0.508
CICD	0.846	0.795	0.572	0.525

Table 2: Ablation analysis of CICD.

means CED is deleted from our model; 3) **-selected I** refers to the selected mechanism is removed from ISI; 4) **-interaction I** represents the co-interaction unit of ISI is replaced by concatenation operation; 5) **-ISI** corresponds to ISI is separated; and 6) **-inconsistency loss** means the inconsistency loss is removed. As shown in Table 2, we observe that:

- The removal of each module (-CED or -ISI) weak-ens the performance of CICD, presenting from 4.2% to 5.5% degradation in micF1, and the strip-ping of different layers (like -selected I and -interaction I) of each module also reduces the model performance, reducing at least 2.4% performance in micF1, which describes the effectiveness of each component and the organic integrity of CICD.

- -CED reflects the lowest performance in all simpli-fied models, decreasing 5.5% and 4.6% in micF1 on the two datasets, respectively, which elaborates the effectiveness of our CICD capturing the collec-tive cognition-view global evidence. Meanwhile, -ISI underperforms CICD, showing 4.3% and 4.2% degradation in micF1 on the two datasets respec-tively, which conveys the necessity of the explo-ration of local key evidence fragments from indi-vidual cognition view.

- When compared with -inconsistency loss, CICD significantly improves the performance on the two datasets with the help of inconsistency loss unit, which verifies the effectiveness of our model rely-

ing on inconsistency loss to discover shared valuable semantics between global and local evidence.

4.3.3 Evaluation of Co-Interaction Networks

To obtain a more detailed understanding of the superiority of our co-interaction networks (CoI), we compare CoI with the following prevalent interaction networks: 1) **MLP** (*Multilayer Perceptron*) acts as an interaction strategy to automatically abstract the integrated representation of claims and articles; 2) **Self-Att** (*Self-attention Networks*) (Vaswani et al., 2017) adopts the claim as query, and relevant articles to serve as values and keys for interaction; 3) **Biaf-Att** (*Biaffine Attention*) (Ma et al., 2019) measures the degree of semantic matching for interaction; and 4) **Symm-Intr** (*Symmetric interaction attention*) (Tao et al., 2019) is exploited to model the interaction between claims and articles. Specifically, we investigate the performance and time cost of these methods on Snopes and PolitiFact based on Linux CentOS with NVIDIA TITAN Xp GPU, as shown in Figure 3. We observe that:

From the overall performance of all methods, our method achieves the optimal performance, outperforming other methods by more than 5.1% and 5.6% performance in micF1, respectively. From the indicator of time cost, our method saves a great deal of time. Compared with Self-Att and Symm-Intr, our method saves from 500 to 1,000 seconds in time cost on the two datasets, respectively. The reason is that the structures of multiple mappings of self-attention networks and the repeat stacks of symmetric attention delay the efficiency. Although the time cost of our method is higher than that of MLP and Biaf-Att, the performance of both methods is unsatisfactory, which is lower than our method al least 2.6% and 3.7% in micF1 on both datasets. On the whole, these adequately manifest the superiority of our method.

4.3.4 Evaluation of Hierarchical Attention Decoder

To verify the effectiveness of the internal structure of hierarchical attention decoder (HAD) in CED, we ablate HAD with the following models: **-word.**, **-sentence.**, and **-merge.** respectively denote HAD removing word-level attention α, sentence-level attention β, and merged semantics γ. **decoder.** represents the vanilla decoder. Experimental results are shown in Table 3, we observe that: first, the removal of any module of HAD could weaken the performance of the model, which confirms the effectiveness of each module. Second, in addition to the basic decoder,

Methods	Snopes		PolitiFact	
	micF1	macF1	micF1	macF1
-word.	0.827	0.784	0.558	0.517
-sentence.	0.815	0.770	0.545	0.502
-merge.	0.833	0.787	0.562	0.520
decoder.	0.804	0.758	0.535	0.489
CICD	0.846	0.795	0.572	0.525

Table 3: Evaluation of hierarchical attention decoder.

Methods	Dev.		Test.	
	Acc.	F1	Acc.	F1
NSMN	0.697	0.431	0.621	0.398
HAN	0.720	0.488	0.669	0.446
GEAR	0.738	0.492	0.708	0.474
KGAT	0.745	0.501	0.716	0.485
CICD (Ours)	**0.763**	**0.525**	**0.731**	**0.497**

Table 4: Results of different baselines on FEVER.

our model achieves the most prominent boost with the support of sentence-level attention, which proves the effectiveness of HAD fusing sentence-level semantics to capture global semantics of HAD.

To further investigate the contribution of sentence-level semantics to the global evidence, we take Figure 1 as an example to visualize the global evidence generated by our model with and without sentence-level attention, respectively. As shown in Figure 4, we observe that the model with sentence-level attention focuses more on the sentences with maximum weight, that is, R4, such as the words 'do not spread' and 'refuted it spreads in the air', while the model without sentence-level attention does not identify which relevant articles are more valuable, so that they concentrate more on R2 and R3, like 'get infected husband' and 'not all types of mosquitoes'. These fully prove the effectiveness of sentence-level semantics for the generation of global evidence.

4.4 Experiments on FEVER

To examine the extensibility of our model, we also compare CICD and the following state-of-the-art baselines on FEVER dataset: 1) **NSMN**: The pipeline-based system, Neural Semantic Matching Network (Nie et al., 2019), conducts document retrieval, sentence selection, and claim verification jointly for fact extraction and verification; 2) **HAN**: It has introduced in Section 4.3.1; 3) **GEAR**: A graph-
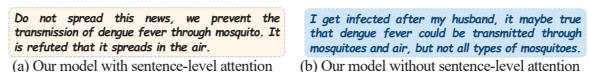

(a) Our model with sentence-level attention (b) Our model without sentence-level attention

Figure 4: The sequences generated by our model with and without sentence-level attention, respectively.
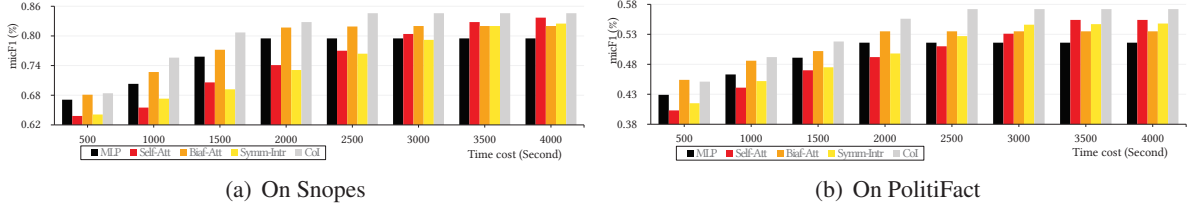
| (a) On Snopes | (b) On PolitiFact |

Figure 3: The performance comparison between co-interaction networks (CoI) and some prevalent interaction networks.

(a) Generated sequence from CED

Claimed 120 million Americans died of coronavirus is a serious error, Biden's mistake, the screenshot is one-sided.

(b) The interpretation via visualization of CICD

Claim: 120 million people had died of the COVID-19 coronavirus disease.
R1: That's what this video screenshot shows "People don't have a job, people don't know where to go, they don't know what to do," Biden said Thursday. "Now we have over 120 million dead from COVID."
R2: Is this a prophecy? Coronavirus is terrible. More than one-third of the country died from it.
R3: He just falsely claimed 120 million Americans died from the coronavirus. "That's a serious error. That's not a permissible type of error," Trump said. Trump took to Twitter to criticize Biden's mistake, which he called, "mortifying stupid."
R4: Joe Biden short circuiting again. You better get programmed better next speech.
R5: The screenshot of the video is one-sided, it is only a segment of the video. However, he corrected himself immediately to say the number was actually 120,000.

(c) The visualization of captured words by attention weights in ISI

Claim: 120 million people had died of the COVID-19 coronavirus disease.
R1: That's what this video screenshot shows "People don't have a job, people don't know where to go, they don't know what to do," Biden said Thursday. "Now we have over 120 million dead from COVID."
R2: Is this a prophecy? Coronavirus is terrible. More than one-third of the country died from it.
R3: He just falsely claimed 120 million Americans died from the coronavirus. "That's a serious error. That's not a permissible type of error," Trump said. Trump took to Twitter to criticize Biden's mistake, which he called, "mortifying stupid."
R4: Joe Biden short circuiting again. You better get programmed better next speech.
R5: The screenshot of the video is one-sided, it is only a segment of the video. However, he corrected himself immediately to say the number was actually 120,000.
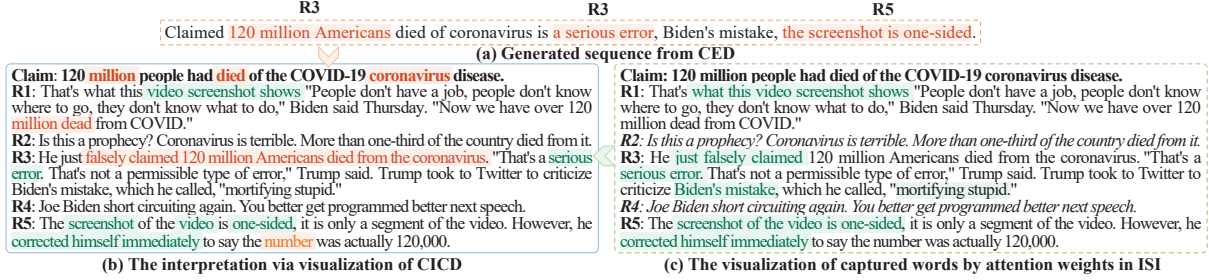
Figure 5: The Interpretability and transparency of different modules of CICD via a false sample on Snopes.

based evidence aggregating and reasoning model (Zhou et al., 2019) enables information to transfer on a fully-connected evidence graph and then utilizes different aggregators to collect multi-evidence information; 4) **KGAT**: Kernel graph attention network (Liu et al., 2020) conducts more fine-grained fact verification with kernel-based attentions, where using BERT (Base) encoder with ESIM retrieved sentences.

As shown in Table 4, we observe that: CICD outperforms the two pipelines (NSMN and HAN) by from 4.3% to 11.0% boost in accuracy, respectively. This is because these two baselines lack the integration and reasoning process between relevant articles when capturing evidence. CICD boosts the performance in comparison with GEAR and KGAT, showing at least 1.8% and 1.5% improvement in accuracy on development and testing sets, respectively. The reason may be that although the two graph-based models aggregate and reason information from relevant articles to collect multi-evidence, they treat each relevant article equally, leading to individual-cognitive relevant articles with some biased semantics interfering with their reasoning process. It is more feasible for our model to discover global evidence and local key evidence fragments comprehensively from the perspectives of collective and individual cognition.

4.5 Case Study: Cognition-view Explanation Analysis

To interpret the results of our model more transparently and intuitively, we visualize the outputs of each module of CICD as shown in Figure 5, where Figure

5 (a) is the sequence generated by CED module, and the highlighted words in Figure 5(b) and 5 (c) are respectively the words captured by CICD to interpret the results and the words obtained by ISI module to obtain the evidence fragments. We could learn:

- ISI ignores some articles with pale and feeble semantics (R2 and R4), and selects the articles with more valuable semantics (R1, R3, and R5) and captures multiple local evidence fragments, such as 'this video screenshot shows' (E1), 'serious error' (E2), and 'screenshot of the video is one-sided' (E3). Particularly, fragment E1 is misleading, which reflects the deviation of individual cognition.

- The sequence generated by CED effectively gains available evidence '120 million Americans a serious error' and 'the screenshot is one-sided' through balancing the possible evidence semantics in relevant articles from a global perspective.

- By constraining global and local evidence, CICD disciplines the misleading evidence fragment E1 captured by ISI, and finally highlights the shared salient evidence between the both as the final interpretability of the verification results.

5 Conclusion

In this paper, we proposed a unified dual-view model based on the perspectives of collective and individual cognition for interpretable claim verification, which constructed collective cognition view-based

encoder-decoder module to generate global evidence and designed individual cognition view-based selected interaction module to explore local key evidence segments. Besides, we introduced inconsistent loss to penalize the disagreement between global and local evidence for promoting the capture of consistent shared evidence. Experiments on three different widely used datasets demonstrated the effectiveness and interpretability of our model. In the future, we plan to expand the work as follows: 1) Developing questioning mechanism to filter the suspicious evidence; and 2) Integrating social cognition, psychology, and other interdisciplinary knowledge to improve the interpretability of claim verification.

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE.

Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539.

N. Boogert, J. Madden, J. Morand-Ferron, and A. Thornton. 2018. Measuring and understanding individual differences in cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nic Fleming. 2020. Coronavirus misinformation, and how scientists can help to fight it. *Nature*, 583:155–156.

Sam Goldstein and Jack A. Naglieri, editors. 2011. *Halo Effect*, pages 725–725. Springer US, Boston, MA.

A. Greenwald, D. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80.

Paul Hoffman. 2018. An individual differences approach to semantic cognition: Divergent effects of age on representation, retrieval and selection. *Scientific Reports*, 8.

J. Ji, E. Holmes, C. MacLeod, and F. Murphy. 2019. Spontaneous cognition in dysphoria: reduced positive bias in imagining the future. *Psychological Research*, 83:817 – 831.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Sumeet Kumar and Kathleen M Carley. 2019. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5047–5058.

Shuaipeng Liu, Shuo Liu, and Lei Ren. 2019. Trust or suspect? an empirical ensemble framework for fake news classification. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 11–15.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *ACL*, pages 2561–2571.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph

representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1165–1174.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Angela K. Troyer. 2011. *Primacy Effect*, pages 2017–2018. New York, NY.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Aiyana K. Willard and A. Norenzayan. 2017. spiritual but not religious: Cognition, schizotypy, and conversion in alternative beliefs. *Cognition*, 165:137–146.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4636–4645.

Lianwei Wu, Yuan Rao, Hao Liang, Ambreen Nazir, et al. 2020a. Dtca: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035.

Lianwei Wu, Yuan Rao, Ambreen Nazir, and Haolin Jin. 2020b. Discovering differential features: Adversarial learning for information credibility evaluation. *Information Sciences*, 516:453–473.

Lianwei Wu, Yuan Rao, Ling Sun, and Wangbo He. 2021. Evidence inference networks for interpretable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1165–1174.

Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2020c. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of IJCAI*.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5644–5651.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367. Springer.