Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification

Lianwei Wu, **Yuan Rao***, **Xiong Yang**, **Wanzhen Wang** and **Ambreen Nazir**
Lab of Social Intelligence and Complexity Data Processing,
School of Software Engineering, Xi'an Jiaotong University, China
Shannxi Joint Key Laboratory for Artifact Intelligence(Sub-Lab of Xi'an Jiaotong University), China
Research Institute of Xi'an Jiaotong University, Shenzhen, China
{stayhungry, youngpanda, kelyin0417, ambreen.nazir}@stu.xjtu.edu.cn, raoyuan@mail.xjtu.edu.cn

Abstract

Exploring evidence from relevant articles to confirm the veracity of claims is a trend towards explainable claim verification. However, most strategies capture the top-k check-worthy articles or salient words as evidence, but this evidence is difficult to focus on the questionable parts of unverified claims. Besides, they utilize relevant articles indiscriminately, ignoring the source credibility of these articles, which may cause quiet a few unreliable articles to interfere with the assessment results. In this paper, we propose Evidence-aware Hierarchical Interactive Attention Networks (EHIAN) by considering the capture of evidence fragments and the fusion of source credibility to explore more credible evidence semantics discussing the questionable parts of claims for explainable claim verification. EHIAN first designs internal interaction layer (IIL) to strengthen deep interaction and matching between claims and relevant articles for obtaining key evidence fragments, and then proposes global inference layer (GIL) that fuses source features of articles and interacts globally with the average semantics of all articles and finally earns the more credible evidence semantics discussing the questionable parts of claims. Experiments on two datasets demonstrate that EHIAN not only achieves the state-of-the-art performance but also secures effective evidence to explain the results.

1 Introduction

There are a large number of unverified claims on social media, such as hoaxes and fake news, which are a widespread menace that has resulted in protests and violence around the globe. Research indicates that fake news accounts for nearly 6% of all news consumption during the US presidential election (2016) [Grinberg *et al.*, 2019], and even some institutions utilize Facebook to selectively expose false claims that affect voters' attitudes [Guess *et al.*, 2018]. Vosoughi et al. [2018] stated that effects were more pronounced for false political news than for false news about terrorism, natural disaster, science, urban legends or financial information. These have driven social media organizations and government institutions to scramble together to tackle this problem.

Currently, the research for claim verification could be divided into three stages: 1) Hoax-debunkers[1][2] are employed to manually assess whether specific claims are true or false, which guarantees high accuracy, but it is difficult to cope with the increasing number of unverified claims; 2) Various methods for automatic claim verification are proposed based on deep neural networks, which learn credibility indicators from the perspectives of semantics [Ma *et al.*, 2018; Wu *et al.*, 2020], emotions [Ajao *et al.*, 2019], write styles [Gröndahl and Asokan, 2019], and stances [Wu *et al.*, 2019] from claims. Although being effective, they cannot explain one claim's correctness in practice; and 3) A recent trend is to discover evidence from relevant articles for explainable claim verification, which designs interactive models to explore the relationships between claims and the articles through semantic conflicts [Popat *et al.*, 2018; Shu *et al.*, 2019; Wu and Rao, 2020], semantic matching [Nie *et al.*, 2019], and semantic entailments [Ma *et al.*, 2019].

However, there are several general drawbacks of the evidence-based methods. **First**, they capture the top-k check-worthy articles or salient words as evidence, lacking focus on the questionable parts of unverified claims, which makes the evidence unable to provide an accurate and effective explanation for the verification results. **Second**, they exploit relevant articles indiscriminately, neglecting the source credibility of these articles, which may lead to the untrustworthy features in unreliable articles interfering with the final results.

To address the above problems, we propose **E**vidence-aware **H**ierarchical **I**nteractive **A**ttention **N**etworks (henceforth, EHIAN) to explore more credible evidence semantics for explainable claim verification. Specifically, in EHIAN, to focus on fragments of evidence from relevant articles, we design an internal interaction layer (IIL) that adopts the combination of self-attention networks and symmetrical attention networks to facilitate full interaction of the claim and each relevant article. In order to earn the more credible evidence semantics discussing the questionable parts of claims, we develop global inference layer (GIL) that devises gated affine transformation to coherently integrate the source features of

[1] Corresponding Author.

[1] https://www.snopes.com

[2] https://www.politifact.com

articles and evidence fragments, and explores global interaction inference to promote the interaction between claims and relevant articles again by combining the average semantics of all articles for mitigating the impact of extreme semantics on result discrimination. Experimental results reveal that EHIAN not only achieves the state-of-the-art performance but also discovers effective evidence for explaining the results. Our contributions are summarized as follows:

- We propose a novel explainable claim verification framework based on hierarchical interactive networks, which not only captures the questionable parts of claims but also wins the more credible evidence semantics aiming at the parts.
- Developed global inference layer focuses on more credible evidence from key fragments through proposed gated affine transformation, and alleviates the impact of extreme voices on the results with the help of global interaction inference.
- We report state-of-the-art performance in two datasets.

2 Related Work

The existing work endeavors to understand the differences between false and true claims from different perspectives, especially in terms of claim content and sources.

Claim Content. In order to reveal linguistic differences between true and false claims, shallow features [Kakol *et al.*, 2017; Potthast *et al.*, 2018], and deep features, like semantic [Ma *et al.*, 2018], stance-based [Dungs *et al.*, 2018], emotional [Giachanou *et al.*, 2019], and stylistic [Gröndahl and Asokan, 2019] have been exploited. As an example, Rashkin et al. [2017] compared the language of true claims with that of satire, hoaxes, and propaganda to find linguistic characteristics of untrustworthy text. Giachanou et al. [2019] incorporated emotional signals extracted from the text of claims to differentiate between credible and non-credible ones. More recently, the methods that explore evidence to enrich the semantics of claims to improve performance are recognized, which develop effective interaction models to discover the relationship between evidence articles and claim content by considering semantic matching [Nie *et al.*, 2019], semantic conflicts [Popat *et al.*, 2018; Shu *et al.*, 2019], and textual entailment [Ma *et al.*, 2019]. Popat et al. [2018] proposed an evidence-aware attention model that aggregates salient words from source news articles as the main evidence for finding false claims. Ma et al. [2019] employed representation learning to embed sentence-level evidence based on textual entailment and natural language inference. In this work, different from the existing evidence-based models only capturing check-worthy sentences or salient words as evidence, the hierarchical interactive model we developed discovers more credible and accurate evidence semantics for catching the questionable parts of claims.

Source Credibility. The source credibility of claims is crucial auxiliary information for claim verification. As false claims are usually published by unbelievable individuals or automatic bots, source credibility plays a crucial role in message communication [Shu *et al.*, 2017]. With the aid of information sources, Popat et al. [2016] found that tweets from highly credible institutions and individuals are mostly correct and then made use of source reliability of articles reporting the claim to assess its credibility and obtain satisfied assessment results. Thus, when information comes from different websites, the credibility of websites represents the credibility of information to a certain extent [Li *et al.*, 2016]. Considering that most evidence-based methods usually ignore the articles' source credibility, we measure and integrate multiple metadata features of article sources for obtaining more credible evidence semantics.

3 Evidence-aware Hierarchical Interactive Attention Networks

As shown in Figure 1, EHIAN consists of a 4-level hierarchical structure: input embedding layer, internal interaction layer, global inference layer, and task learning layer.

3.1 Input Embedding Layer

The inputs of EHIAN include a claim sequence and n relevant article sequences. For any sequence containing l tokens, it can be expressed as $X = \{x_1, x_2, ..., x_l\}$, $X \in \mathbb{R}^{d\times l}$, where each token $x_i \in \mathbb{R}^d$ is a d-dimensional vector obtained by pre-trained BERT model [Devlin *et al.*, 2019]. Here, the embeddings of one claim and n relevant articles are respectively represented as X^c,X^a_1, ..., and X^a_n. Particularly, we use the average embeddings of all relevant articles as a newly-synthesized relevant article X^a_{avg} (the right part in Figure 1), which serves as the semantic benchmark of relevant articles to attend to Global Interaction Inference (Section 3.3) for alleviating the interference of extreme semantics in individual articles.

3.2 Internal Interaction Layer (IIL)

In order to capture key evidence fragments, we design internal interaction layer (IIL) consisting of two self-attention modules and a symmetrical interaction module to enable deep interaction and matching between claims and relevant articles.

Self-attention Module. We adopt multi-head self-attention mechanism [Vaswani *et al.*, 2017] to explicitly learn the dependencies between any two characters in sequence and capture the inner structure information of sequence. Given query matrix $Q \in \mathbb{R}^{l\times d}$, key matrix $K \in \mathbb{R}^{l\times d}$, and value matrix $V \in \mathbb{R}^{l\times d}$, the scaled dot-product attention is described as

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \qquad (1)$$

where we set $Q = K = V = X^c$ in the claim, and $Q = K = V = X^a_i$ in the relevant article.

Multi-head attention first linearly projects the queries, keys, and values h times by utilizing different linear projections. Then h results perform the scaled dot-product attention in parallel. Finally, the results of attention are concatenated and once again projected to get the new representation. Formally, the multi-head attention could be formulated as:

$$\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \qquad (2)$$

$$\begin{aligned} O &= \text{MultiHead}(Q, K, V) \\ &= \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^o \end{aligned} \qquad (3)$$

where W^Q_i, W^K_i, W^V_i (all $\in \mathbb{R}^{d\times H}$), and $W^o \in \mathbb{R}^{d\times d}$ are trainable parameters and H is d/h. O=O^c and O=O^a_i are the outputs aiming at the claim and the relevant article i respectively.

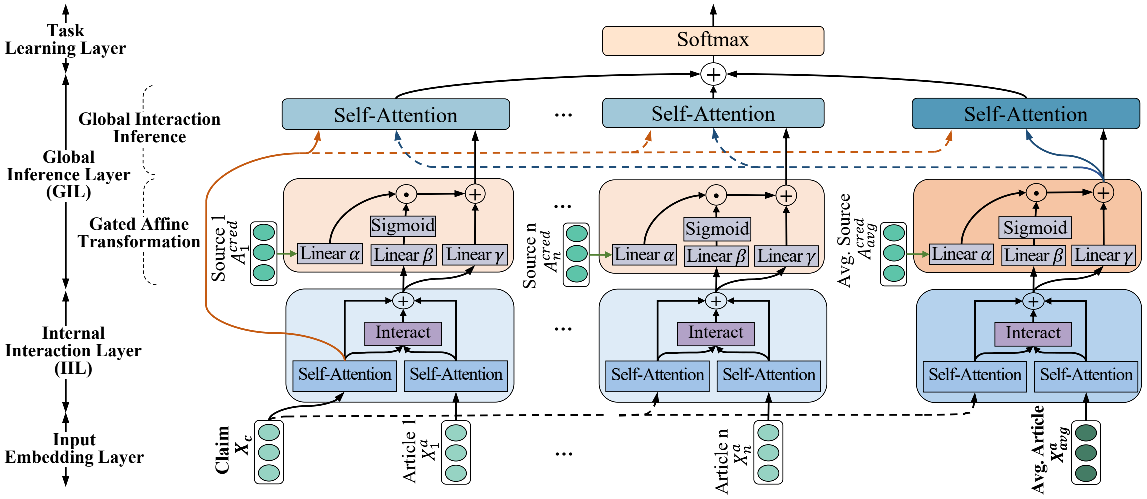


Figure 1: The architecture of EHIAN. The model utilizing organically semantic and source features focuses on the interaction and inference between claims and relevant articles by the following layers: input embedding layer, internal interaction layer, global inference layer, and task learning layer. Particularly, global inference layer consists of two components, i.e., gated affine transformation and global interaction inference.

Symmetrical Interaction Module. We design a symmetrical attention mechanism to boost in-depth interaction and matching between the claim O^c and each relevant article O_i^a. The process can be formalized as follows:

$$I_i = S(O^c, O_i^a)O_i^a \tag{4}$$

$$S(O^c, O_i^a) = \text{softmax}(f(QW)Df(KW)^T) \tag{5}$$

where $S(\cdot)$ is the symmetric function [Huang *et al.*, 2018], f is an RELU activation function. D is a diagonal matrix, where D and W (both $\in \mathbb{R}^{d\times d}$) are parameters. Intuitively, each element of O_i^a is weighted by an importance score defined by the similarity of an element of the claim O^c and that of the relevant article O_i^a. The result I_i is the interactive semantics.

Integration. Finally, we integrate the long-term dependencies of the claim and the article, and interactive semantics between the claim and the relevant article to gain evidence fragments.

$$E_i = [O^c; I_i; O_i^a] \tag{6}$$

where ; denotes the concatenation operation.

3.3 Global Inference Layer (GIL)

In order to discover the more credible evidence semantics discussing the questionable parts of claims, we design GIL including two components, i.e., gated affine transformation for fusing the source features of relevant articles to infer more credible evidence semantics, and global interaction inference for mitigating the impact of extreme semantics on results.

Gated Affine Transformation. Following "the more credible the information source is, the more authentic the information will be", we utilize source credibility of relevant articles to reflect indirectly their credibility. Specifically, we enrich source credibility by extracting meta-data features of source websites of relevant articles from Alexa[3] based on websites' domains (due to the original datasets do not include the details of these websites), where these meta-data features involve the following items from the perspectives of authority, activity, and popularity: 1)**Authority**: domain suffix, sites linking in, visited just before, and visited right after; 2) **Activity**: daily pageviews per visitor, Alexa rank 90 day trend, search traffic, bounce rate, and daily time on site; and 3) **Popularity**: site's audience interests, interest level, and top keywords by traffic. Then we quantify them and finally form a 300-dimensional credibility vector A_i^{cred}.

To deeply integrate the source credibility of relevant articles with their semantics, instead of the traditional concatenation, we elaborately devise gated affine transformation to incorporate the credibility vector A_i^{cred} into the outputs of IIL aiming at the article i. Specifically, we first employ linear transformation to map the credibility vector A_i^{cred} and the interactive semantics E_i to obtain the mapping credibility vector $\alpha(A_i^{cred})$, scaling vector $\beta(E_i)$, and shifting vector $\gamma(E_i)$, respectively. Then, a gate mechanism with a sigmoid function $\sigma(\cdot)$, generates a mask-vector from the scaling vector with values between 0 and 1 to select credibility semantics. Finally, the shifting vector adjusts slightly the credibility semantics to achieve the appropriate fusion. Formally, the process can be formalized as follows:

$$\alpha(A_i^{cred}) = W_\alpha E_i + b_\alpha \tag{7}$$

$$\beta(E_i) = W_\beta E_i + b_\beta \tag{8}$$

$$\gamma(E_i) = W_\gamma E_i + b_\gamma \tag{9}$$

$$E_i^{cred} = f(A_i^{cred}, E_i) = \sigma(\beta(E_i)) \odot \alpha(A_i^{cred}) + \gamma(E_i) \tag{10}$$

where W_α, W_β, W_γ, b_α, b_β, b_γ are learnable parameters and $\odot$ denotes element-wise multiplication. Especially, the integration of the average credibility vector E_{avg}^{cred} of all source vectors is the same as the single relevant article's, E_{avg}^{cred} reflects the overall credibility of relevant articles under the claim.

Global Interaction Inference. To balance and mitigate the interference of extreme semantics in some relevant articles, we introduce the average integration E_{avg}^{cred} of all relevant articles into the interaction again between the claim and each

[3] https://www.alexa.com/siteinfo

relevant article. Specifically, we adopt self-attention networks to achieve the interaction of the three types of semantic features, where the structure of networks is same as self-attention module in IIL. Here, we adopt the outputs O^c of self-attention module in IIL at claims as query Q, key K is the outputs E_i^{cred} of gated affine transformation concerning each relevant article, and value V is the average interaction E_{avg}^{cred} about average relevant articles. Finally, we adopt concatenation to integrate all outputs of GIL.

$$F = [G_1; G_2; ...; G_n; G_{all}] \tag{11}$$

where G_i denotes the outputs of global inference layer of single relevant article i and G_{all} means the outputs of global inference layer of all relevant articles. At this time, F contains the highly trusted evidence semantics discussing the questionable parts of claim.

3.4 Task Learning Layer

As the last layer, softmax function first emits the prediction of probability distribution for task learning. Then, a global loss function forces the model to minimize the cross-entropy error for a training sample with ground-truth label y:

$$p = \mathrm{softmax}(W_p F + b_p) \tag{12}$$

$$\mathrm{Loss} = -\sum ylogp \tag{13}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We use two public fact-checking datasets, i.e., Snopes and PolitiFact provided by Popat et al. [2018] for evaluation. **Content.** Both contain news claims along with their credibility labels, sets of relevant articles, and their respective web source domains. **Labels.** Snopes is labeled as *true* and *false* while PolitiFact is originally assigned six credibility labels: true, mostly true, half true, mostly false, false, and pants on fire. Like Ma et al. [2019], we merge the six labels into *true*, *false*, and *mixed*. In detail, mostly true, half true, and mostly false are consolidated as *mixed*, and treat false and pants on fire as *false*. **Distribution.** Snopes and PolitiFact include 4,341 and 3,568 news claims, and 29,242 and 29,556 relevant articles that both include respectively a total of 336 web sources.

Additionally, we adopt micro-/macro-averaged F1, class-specific precision (P), recall (R), and F-measure (F1) as evaluation metrics. We hold out 10% of all data as validation data for parameter tuning, and conduct 5-fold cross-validation on the remaining 90% of the data.

4.2 Settings

For parameter configurations, we apply the pre-trained BERT-base model [Devlin *et al.*, 2019] to initialize our token embeddings and their length is set to 300. Due to no parameter depends on the number of relevant articles n, instead of intercepting a fixed number, we set n to vary with claims in datasets. In self-attention networks, attention heads and blocks are set to 6 and 2 respectively, and the dropout of multi-head attention is set to 0.7. Moreover, all the models are trained to use Adam optimizer [Kingma and Ba, 2014] with a learning rate of 0.002 and mini-batches of size 64 to minimize categorical cross-entropy loss. We employ L2-regularizers with the fully connected layer. Also, the dropout is 0.5.

4.3 Performance Comparison

We compare EHIAN and several state-of-the-art baselines.

- **SVM** detects fake news relying on manually extracted features (e.g., bag-of-words, ngrams, etc.) from relevant articles [Thorne and Vlachos, 2018].
- **CNN** [Wang, 2017] captures semantics by different convolutional window sizes for fake news detection. Here, we only use claim content without considering meta-data features.
- **LSTM** [Rashkin *et al.*, 2017] takes the word sequence of a claim as the input and predicts credibility rating.
- **DeClarE** [Popat *et al.*, 2018] presents attention networks to aggregate signals like salient words from external relevant articles and their sources for claim verification.
- **HAN** [Ma *et al.*, 2019] focuses on learning coherent evidence as well as their semantic relatedness with the claim.
- **HAN-ba** [Ma *et al.*, 2019] is a variant of HAN that uses the biaffine attention to replace gated attention as the coherence component.

We implement our model with Tensorflow[4]. Due to DeClarE is not open-source, we adopt the results provided by the experiments of Ma et al. [2019] with Theano[5]. Other baselines are implemented through their source codes.

Table 1 shows the experimental results and we observe that:

- CNN and LSTM only using content features are comparable with SVM incorporating many handcrafted features, which reveals neural networks indeed help to learn better hidden representation. DeClarE outperforms CNN and LSTM because it not only learns deep features via the neural model but also focuses on the differential features like salient words by attention networks. HAN and HAN-ba achieve better performance than DeClarE, which consider the relationship like coherence and textual entailment between relevant articles.
- Our model obtains competitive performance than other baselines. Unlike HAN and HAN-ba that treat all relevant articles equally, our model takes the credibility of relevant articles into account to strengthen high-reliability evidence semantics. Unlike DeClarE only capturing salient words as evidence, our model learns evidence semantics discussing the questionable parts of claims through hierarchical interaction between claims and articles.

4.4 Discussion

Ablation Analysis

To evaluate the effectiveness of different components of EHIAN, we ablate EHIAN into the following simplified models: **-IIL** denotes that EHIAN removes IIL. **-Interact** means EHIAN replaces symmetrical interaction module in IIL with concatenation. We remove GIL from EHIAN as **-GIL**. **-Affine** means gated affine transformation is replaced by concatenation in EHIAN. **-Avg.** means that the average semantics of all relevant articles are no longer the values of self-attention networks in GIL, but are replaced by the outputs of each gated affine transformation.

As shown in Table 2, we have the following observations:

[4] https://www.tensorflow.org/

[5] http://deeplearning.net/software/theano

Methods	Snopes								PolitiFact				
			True			False					True	False	Mixed
	micF1	macF1	P	R	F1	P	R	F1	micF1	macF1	F1	F1	F1
SVM	0.704	0.649	0.459	0.584	0.511	0.832	0.747	0.786	0.450	0.421	0.440	0.547	0.277
CNN	0.721	0.636	0.477	0.440	0.460	0.802	0.822	0.812	0.453	0.402	0.368	0.566	0.270
LSTM	0.689	0.642	0.441	0.512	0.517	0.834	0.716	0.771	0.463	0.413	0.452	0.561	0.228
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
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
Ours	**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**

Table 1: Performance comparison of EHIAN against the baselines on Snopes and PolitiFact datasets.

Methods	Snopes								PolitiFact				
			True			False					True	False	Mixed
	micF1	macF1	P	R	F1	P	R	F1	micF1	macF1	F1	F1	F1
-IIL	0.760	0.702	0.551	0.621	0.584	0.845	0.834	0.839	0.497	0.462	0.466	0.597	0.321
-GIL	0.741	0.683	0.534	0.598	0.564	0.834	0.826	0.830	0.486	0.455	0.458	0.581	0.311
-Interact	0.807	0.761	0.602	0.758	0.671	0.871	0.868	0.869	0.525	0.492	0.494	0.630	0.344
-Affine	0.802	0.757	0.600	0.746	0.665	0.865	0.857	0.861	0.534	0.486	0.491	0.632	0.347
-Avg.	0.810	0.761	0.603	0.754	0.670	0.869	0.862	0.865	0.529	0.481	0.485	0.628	0.339
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

Table 2: Results of ablation test of our EHIAN on Snopes and PolitiFact datasets.

- **Effectiveness of internal interaction layer.** EHIAN boosts about 7.1% and 5.7% in micF1 on Snopes and PolitiFact respectively compared with -IIL, which indicates the interaction between claims and articles via IIL is effective.
- **Effectiveness of global inference layer.** When compared with -GIL, EHIAN significantly improves performance with the help of GIL, showing 9% and 6.8% boost in micF1 on the two datasets respectively, which explains that the effectiveness of GIL to obtain more credible evidence semantics discussing the questionable parts of claims.
- **Effectiveness of symmetrical interaction module.** Analysis of the results of -Interact and EHIAN, EHIAN obtains the better performance relying on the interaction module, which illustrates that the effectiveness of EHIAN using symmetrical attention mechanism to achieve the interaction.
- **Effectiveness of gated affine transformation.** By introducing gated affine transformation, EHIAN improves the performance as compared with -Affine, showing 2.7% and 2.3% improvement in macF1 on the two datasets, respectively. It proves the effectiveness of gated affine transformation fusing meta data and semantics.
- **Effectiveness of global interaction inference.** EHIAN consistently outperforms -Avg., presenting 1.3% and 1.8% boost in macF1 on the two datasets, respectively, which confirms that EHIAN designing global interaction inference module to mitigate extreme bias voices is effective.

Evaluation of Gated Affine Transformation

To further evaluate the superiority of the fusion mode for gated affine transformation (i.e., gate+affine) in GIL, we conduct experiments to compare with the following strategies: **conc.:** Semantics and meta data (a.k.a. the credibility vector) are fused through concatenation as a baseline. **add.:** We use additions to integrate meta data and semantics as a baseline. **heuristics:** Matching heuristics [Mou *et al.*, 2016] is used to fuse meta data and semantics. **affine:** Attentional affine transformation [Margatina *et al.*, 2019] replaces gate+affine to incorporate meta data and semantics. **gate:** We adopt attentional feature-based gating [Margatina *et al.*, 2019] to combine meta data and semantics. **gate+emb.conc.:** We utilize the combination of attentional feature-based gating and attentional concatenation [Margatina *et al.*, 2019] to fuse meta data and semantics. The experimental results are illustrated in Table 3, we observe that:

- Compared with two baseline strategies, the improved methods achieve varying degrees of boost (up to 2.7% improvement in micF1) on the two datasets, where gate mechanism relying on filtering features gains better performance than heuristics and affine. gate+emb.conc. taking a combined way reflects the best performance in all improved methods.
- Our method outperforms the improved methods on the two datasets, showing from 1.1% to 1.4% boost in macF1 in comparison with gate+emb.conc., which illustrates the superiority of our gated affine transformation.

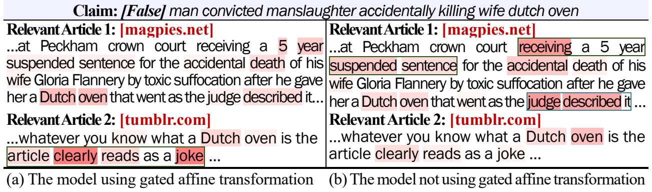
4.5 Case Study

The Visualization of Features Learned from GIL

We examine the capture of valuable features in two components of GIL. In the first component, we respectively map the outputs of both gated affine transformation and IIL (not using meta-data features of source websites) to the input elements

	Methods	Snopes		PolitiFact	
		micF1	macF1	micF1	macF1
Baselines	conc.	0.790	0.755	0.520	0.475
	add.	0.793	0.751	0.524	0.477
The improved methods	heuristics	0.802	0.758	0.529	0.481
	affine	0.804	0.760	0.527	0.483
	gate	0.812	0.766	0.536	0.489
	gate+emb.conc.	0.820	0.773	0.540	0.495
Our method	gate+affine	0.831	0.784	0.554	0.509

Table 3: Comparison between our gated affine transformation and existing fusion strategies on Snopes and PolitiFact.



Claim: ***[False]*** *man convicted manslaughter accidentally killing wife dutch oven*

Relevant Article 1: [magpies.net]
...at Peckham crown court receiving a 5 year suspended sentence for the accidental death of his wife Gloria Flannery by toxic suffocation after he gave her a Dutch oven that went as the judge described it...

Relevant Article 2: [tumblr.com]
...whatever you know what a Dutch oven is the article clearly reads as a joke ...

(a) The model using gated affine transformation

Relevant Article 1: [magpies.net]
...at Peckham crown court receiving a 5 year suspended sentence for the accidental death of his wife Gloria Flannery by toxic suffocation after he gave her a Dutch oven that went as the judge described it ...

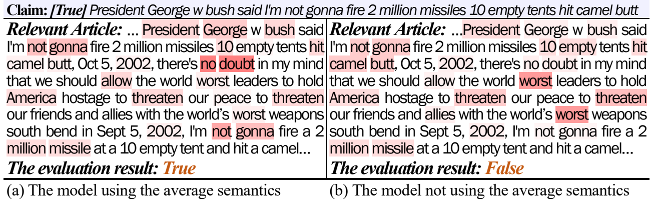
Relevant Article 2: [tumblr.com]
...whatever you know what a Dutch oven is the article clearly reads as a joke ...

(b) The model not using gated affine transformation

Figure 2: Interpretation via visualization of gated affine transformation in GIL. [magpies.net/tumblr.com] denotes the source websites.

of word-level. The visualized results are described in Figure 2, where the boxes are the captured evidence fragments. We observe that the model integrating meta-data features more focuses on the evidence fragments in the website with higher credibility ('tumblr.com'), i.e., 'article clearly reads as a joke', and accurately capture the evidence semantics 'clearly' and 'joke'. Conversely, the model not relying on meta-data features equally captures the keywords from relevant articles in different credibility websites, which misleads the model to obtain unreliable evidence fragments from low credibility website ('magpies.net'), resulting in some non-evidence salient words being captured, like 'receiving', and 'judge described'.

Additionally, in the same way, we respectively get the visualization whether the model integrates the average integration E^{cred}_{avg} in global interaction inference, as shown in Figure 3. We observe that the integrated model captures more accurate keywords such as 'not gonna' and 'no doubt', which is not affected by some extreme words, like 'worst' and 'threaten', and finally determines that the claim is true. Conversely, the model not integrated does not consider the global semantics of all relevant articles so that some extreme words are captured, which ultimately misleads the evaluation result (i.e., false).



Claim: ***[True]*** *President George w bush said I'm not gonna fire 2 million missiles 10 empty tents hit camel butt*

Relevant Article: ... President George w bush said I'm not gonna fire 2 million missiles 10 empty tents hit camel butt, Oct 5, 2002, there's no doubt in my mind that we should allow the world worst leaders to hold America hostage to threaten our peace to threaten our friends and allies with the world's worst weapons south bend in Sept 5, 2002, I'm not gonna fire a 2 million missile at a 10 empty tent and hit a camel...

The evaluation result: ***True***

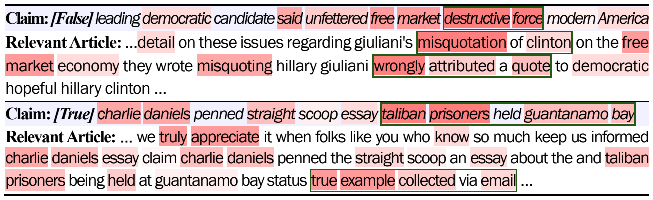
(a) The model using the average semantics

Relevant Article: ...President George w bush said I'm not gonna fire 2 million missiles 10 empty tents hit camel butt, Oct 5, 2002, there's no doubt in my mind that we should allow the world worst leaders to hold America hostage to threaten our peace to threaten our friends and allies with the world's worst weapons south bend in Sept 5, 2002, I'm not gonna fire a 2 million missile at a 10 empty tent and hit a camel...

The evaluation result: ***False***

(b) The model not using the average semantics

Figure 3: Interpretation via visualization of global interaction inference.



Claim: ***[False]*** *leading democratic candidate said unfettered free market destructive force modern America*

Relevant Article: ...detail on these issues regarding giuliani's misquotation of clinton on the free market economy they wrote misquoting hillary giuliani wrongly attributed a quote to democratic hopeful hillary clinton ...

Claim: ***[True]*** *charlie daniels penned straight scoop essay taliban prisoners held guantanamo bay*

Relevant Article: ... we truly appreciate it when folks like you who know so much keep us informed charlie daniels essay claim charlie daniels penned the straight scoop an essay about the and taliban prisoners being held at guantanamo bay status true example collected via email ...

Figure 4: Interpretation via visualization of attention weights in EHIAN. [True/False] indicates the labels of claims.

The Visualization of Features Learned from EHIAN

We visualize the features finally learned from EHIAN. Specifically, we first look up these elements with the largest values from the entire outputs of GIL, and then these elements are mapped into the corresponding values in input embeddings so that we are capable of finding the specific tokens. The visualization results are depicted in Figure 4. We observe that EHIAN gives more attention to the words that are related to the claim, like 'free market' and 'taliban prisoners'. More importantly, we also observe that: **Explainability.** In claims, EHIAN captures the queationable parts of claims with high probability, such as 'destructive force' and 'taliban prisoners held guantanamo bay'. In relevant articles, EHIAN can also capture the key evidence fragments that discusses the questionable parts of claims, e.g., 'misquotation of clinton' and 'wrongly attributed a quote' related to false claim, 'true example collected via email' related to true claim, which is conducive to explaining the verification results of our model.

5 Conclusion

In this paper, we propose evidence-aware hierarchical interactive attention networks (EHIAN) to explore more credible evidence discussing the questionable parts of claims for explainable claim verification. EHIAN first strengthens the interaction between claims and relevant articles to discover key evidence fragments, and then incorporates source features of articles and mitigates the interference of extreme semantics to explore more credible evidence discussing the questionable parts of claims. Experiments on two datasets confirm the effectiveness and interpretability of EHIAN. In the future, we plan to expand the work by enhancing inference to discover conflicts between claims and relevant articles.

Acknowledgments

The research work was supported by National Key Research and Development Program in China (2019YFB2102300); The World-Class Universities (Disciplines) and the Characteristic Development Guidance Funds for the Central Universities of China (PY3A022); Ministry of Education Fund Projects (18JZD022 and 2017B00030); Shenzhen Science and Technology Project (JCYJ20180306170836595); Basic Scientific Research Operating Expenses of Central Universities (ZDYF2017006); Xi'an Navinfo Corp.& Engineering Center of Xi'an Intelligence Spatial-temporal Data Analysis Project (C2020103); Beilin District of Xi'an Science & Technology Project (GX1803). We would like to thank the anonymous reviewers for their constructive comments.

References

[Ajao *et al.*, 2019] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Sentiment aware fake news detection on online social networks. In *ICASSP*, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[Dungs *et al.*, 2018] Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. Can rumour stance alone predict veracity? In *COLING*, pages 3360–3370, 2018.

[Giachanou *et al.*, 2019] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. Leveraging emotional signals for credibility detection. In *SIGIR*, pages 877–880. ACM, 2019.

[Grinberg *et al.*, 2019] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[Gröndahl and Asokan, 2019] Tommi Gröndahl and N Asokan. Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys (CSUR)*, 52(3):45, 2019.

[Guess *et al.*, 2018] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council*, 2018.

[Huang *et al.*, 2018] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *ICLR*, 2018.

[Kakol *et al.*, 2017] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 2017.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2016] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD*, 17(2):1–16, 2016.

[Ma *et al.*, 2018] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. In *ACL*, pages 1980–1989, 2018.

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

[Margatina *et al.*, 2019] Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. Attention-based conditioning methods for external knowledge integration. In *ACL*, 2019.

[Mou *et al.*, 2016] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *ACL*, pages 130–136, 2016.

[Nie *et al.*, 2019] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*, 2019.

[Popat *et al.*, 2016] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *CIKM*, pages 2173–2178. ACM, 2016.

[Popat *et al.*, 2018] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*, pages 22–32, 2018.

[Potthast *et al.*, 2018] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *ACL*, pages 231–240, 2018.

[Rashkin *et al.*, 2017] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937, 2017.

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD*, 19(1):22–36, 2017.

[Shu *et al.*, 2019] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. 2019.

[Thorne and Vlachos, 2018] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *COLING*, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Wang, 2017] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL*, pages 422–426, 2017.

[Wu and Rao, 2020] Lianwei Wu and Yuan Rao. Adaptive interaction fusion networks for fake news detection, 2020. arXiv preprint arXiv:2004.10009.

[Wu *et al.*, 2019] Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. In *EMNLP-IJCNLP*, pages 4636–4645, 2019.

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