PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

Evidence Inference Networks for Interpretable Claim Verification

Lianwei Wu, Yuan Rao, Ling Sun, Wangbo He
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, sunling}@stu.xjtu.edu.cn, raoyuan@mail.xjtu.edu.cn, 744758858@qq.com

Abstract

Existing approaches construct appropriate interaction models to explore semantic conflicts between claims and relevant articles, which provides practical solutions for interpretable claim verification. However, these conflicts are not necessarily all about questioning the false part of claims, which makes considerable semantic conflicts difficult to be used as evidence to explain the results of claim verification, especially those that cannot identify the core semantics of claims. In this paper, we propose evidence inference networks (EVIN), which focus on the conflicts questioning the core semantics of claims and serve as evidence for interpretable claim verification. Specifically, EVIN first captures the core semantic segments of claims and the users' principal opinions in relevant articles. Then, it finely-grained identifies the semantic conflicts contained in each relevant article from these opinions. Finally, EVIN constructs coherence modeling to match the conflicts that queries the core semantic fragments of claims as explainable evidence. Experiments on two widely used datasets demonstrate that EVIN not only achieves satisfactory performance but also provides explainable evidence for end-users.

1 Introduction

The explosive growth of false claims and their erosion on democracy, justice, and public trust greatly aggrandize the demand for false claim verification. Especially with the popularity of social media, the low-cost, large-scale production and dissemination of false claims by malicious rumormongers are unprecedented. Research indicates that although false claims account for only 1% of total news consumption on all platforms (Allen et al. 2020), while the proportion of false claims on social media is more than 6% of total tweets. Since the 2016 U.S. presidential election campaign (Grinberg et al. 2019), false claims have dominated the news cycle, where the top twenty frequently discussed false election stories generated 8,711,000 shares, reactions, and comments on Facebook (Silverman 2016). It has become quite critical to detect the credibility of unverified claims on social media in time to block their spread and refute them.

Claim verification is a tough and challenging task for industry and academia. Previous studies (Castillo, Mendoza, and Poblete 2011; Jin et al. 2016) are devoted to extracting various linguistic and hand-crafted semantic features for differentiating claims, which achieves excellent performance. But it is difficult for these approaches to provide appropriate explanations for the verification results of claims, i.e., why claims are false. Thus, recent research has begun to focus on interpretable claim verification, which generally establishes interactive models to quest the difference (a.k.a. semantic conflicts) between claims and their relevant articles (or comments) as evidence to explain the verified results. Specifically, Popat et al. (2018) devise joint interactions between claims and their web sources to collect salient conflict words as word-level evidence for explaining the correctness of claims. Ma et al. (2019) propose interactive attention networks based on natural language inference (NLI) to learn the sequence consistency, so as to obtain semantic conflicts among relevant articles as sentence-level evidence to verify claims. Wu et al. (2020) construct an interaction fusion network to survey the emotional and semantic conflicts between news and its comments as multi-perspective evidence for enhancing the capability of verification results.

These interaction models acquire the semantic conflicts between claims and relevant articles to explain the final verification results, which is a feasible interpretable idea. The reason is that: different relevant articles, as the opinions of different users on a specific claim, are usually prone to call into question on the wrong parts of the false claims (similar to crowdsourcing rumor refutation), even if they are sometimes incapable of providing an exact conclusion. These query voices are rich in conflicting semantics that may include the reasons for the wrong parts of claims. In consequence, capturing semantic conflicts contributes to strengthening in-

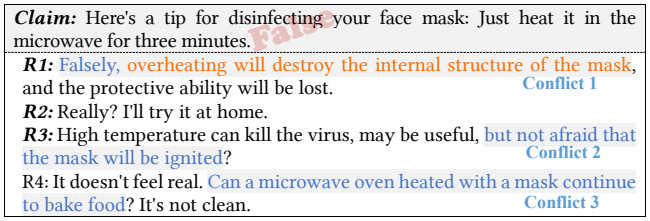


Figure 1: A false claim from Twitter. Among the three semantic conflicts, only conflict 1 is able to provide an effective explanation for the falsity of claim.

terpretability. However, these approaches ignore the fact that the acquired semantic conflicts are not necessarily all about questioning the false part of claims, which makes considerable semantic conflicts difficult to be used as evidence to explain the verification results, especially those that cannot identify the core semantics of claims. Take a concrete example, as shown in Figure 1, there are three conflicts. Only conflict 1 could be taken as a valid piece of evidence which is able to reveal the false part of false news due to it effectively questions the key semantics (i.e., 'just heat') of the claim. But conflict 2 focuses on the flammability of the mask, and conflict 3 considers whether the microwave continue to be used. Neither of them discusses the key wrong parts of the claim and debunks it. Therefore, how to capture the conflicts that focus on the core semantics of claims is a critical problem for enhancing the interpretability of claim verification.

To deal with the above issues, we propose **EV**idence **I**nference **N**etworks (henceforth, EVIN) for interpretable claim verification, which strives to capture the conflicts that focuses on the core fragments of claims as interpretable evidence. Specifically, in EVIN, we first design co-interactive shared layer to enable claims to interact with relevant articles for adaptively capturing the core semantic segments widely concerned by users in claims, as well as the users' holistic opinions in relevant articles, respectively. Then, we design a fine-grained conflict discovery layer that allows the holistic opinions to interact with the individual opinion of each relevant article for exploring the potential semantic conflicts. Finally, to select the conflicts that can be the real evidence, we present evidence-aware coherent layer to construct coherence modeling between the core semantic segments of claims and the obtained conflicts, which matches the conflicts that revolve around the core semantics of claims.

Experiments on two real-world competitive datasets demonstrate that our model outperforms several state-of-the-art approaches by 3.0% and 3.4% points in terms of micF1. Moreover, experimental analysis strongly confirms the interpretability of our model to end-users. To sum up, our main contributions are as follows:

- A novel and refined interpretable claim verification framework (EVIN) is explored, which can distill the conflicts that question the core semantics of claims to act as evidence for explaining the verified results.
- Co-interactive shared layer relying on gate affine absorption module has ability to adaptively focus on the core semantic segments of claims and the holistic opinions of relevant articles, respectively. And evidence-aware coherence layer based on coherence modeling effectively matches the conflicts concentrated on the questioned core semantics of claims.
- Experimental results reveal that our model achieves superior performance on two widely-used datasets. The implementation for our model is publicly available[1].

[1]at the attachment

2 Related Work

In recent years, the task of claim verification on social media has attracted considerable attention. In addition to manual feature extraction for claim verification, existing work has gone through the following two stages.

2.1 Automatic Claim Verification

The methods for automatic claim verification basically rely on neural networks to automatically capture numerous features around the perspective of claim content and its social context, and construct effective classification models for verification. The content-based methods mostly learn the features of n-grams (Wang 2017), semantics (Khattar et al. 2019), emotions (Ajao, Bhowmik, and Zargari 2019), stances (Ma, Gao, and Wong 2018), and writing styles (Gröndahl and Asokan 2019) from claim text. For concrete examples, Karimi et al. (2019) capture style features based on content structure at various language levels, like discourse level by employing rhetorical structure theory for claim verification. Zhou et al. (2020) study news content at four levels: lexicon, syntax, semantic, and discourse, and conduct a supervised machine learning framework to explore potential fake news patterns. Besides, auxiliary information around social context has also been widely investigated. The context-based methods put great emphasis on collecting user profile-based (Shu et al. 2019), propagation structure-based (Wei, Xu, and Mao 2019), comment-based (Ma et al. 2020), source-based (Pennycook and Rand 2019), and platform-based features (Shu, Wang, and Liu 2019). For instance, Zhou et al. (2019) learn propagation network patterns from multiple aspects, i.e., node, ego, triad, community, and the overall network to detect fake news. These methods avoid the heavy labor of the methods based on manual feature extraction, and deeply learn high-level feature representations, which effectively improve the accuracy of the task.

2.2 Interpretable Claim Verification

Due to the methods for automatic claim verification are difficult to provide reasonable explanations for the verification results, the demand for interpretable claim verification is ever-growing, which aims at presenting end-users with reasons to refute rumors through revealing the wrong parts of claims. In this task, existing methods explore semantic conflicts between claims and relevant articles by means of establishing different interactive models to explain verification results. The interactive models could be divided into attention-based interaction models (Popat et al. 2018), gate fusion interactive models (Wu and Rao 2020), and coherence modeling interactive models (Ma et al. 2019), and graph-aware interaction models (Lu and Li 2020). The granularity of captured semantic conflicts involves word-level (Popat et al. 2018), fragment-level (Lu and Li 2020), sentence-level (Ma et al. 2019), and multi-feature (Wu and Rao 2020) conflicts. These methods, which employ semantic conflicts to verify claims, reflect a certain degree of interpretability. But not all conflicts can be used as valid evidence to reasonably explain the results, and they also include considerable conflicts unrelated to claims or even interfere with the verified results. Therefore, we struggle to purify available evidence from semantic conflicts for interpretable claim verification.

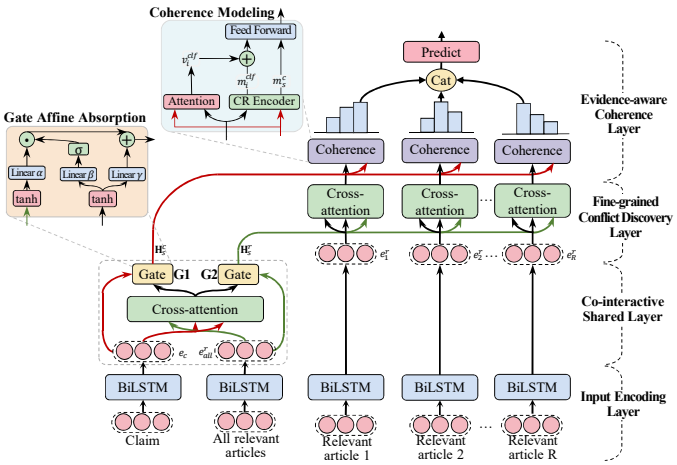


Figure 2: The architecture of our EVIN model.

3 The Proposed Method

In this section, we propose evidence inference networks (EVIN) for interpretable claim verification. The core idea of EVIN is to strengthen the conflicts that question the core semantics of claims and serve as explainable evidence. EVIN consists of 4-level hierarchical structure, input encoding layer, co-interactive shared layer, fine-grained conflict discovery layer, and evidence-aware coherence layer. As shown in Figure 2, we describe each level of EVIN in detail.

3.1 Input Encoding Layer

The inputs of EVIN include three types of sequences: a claim sequence, the concatenated sequence of all relevant articles with quantity R, and the sequence of each relevant article. For any sequence with k tokens, it can be expressed as $\mathbf{X} = \{x_1, x_2, ..., x_k\}$, each $x_t \in \mathbb{R}^d$ is d-dimensional vector which could be initialized with pre-trained word embeddings (Devlin et al. 2019). For the encoding of each sequence $\mathbf{X}_i$, we utilize BiLSTM to learn its sequence features, and adopt the final step's hidden vector $e_i \in \mathbb{R}^{k\times 2h}$ to represent it, where h is the size of hidden units of LSTM. Here, the encodings of the claim, that of the sequence of all relevant articles, and that of the j-th relevant article are represented as e_c, e^r_{all}, and e^r_j $(1 \le j \le R)$, respectively.

3.2 Co-interactive Shared Layer

In order to enable the model to respectively focus on the core semantic segments of claims that are widely concerned by users, as well as the users' holistic opinions discussed in all relevant articles, we design co-interactive shared layer composed of a cross-attention module and two gate affine absorption modules to make claims e_c and all relevant articles e^r_{all} interact with each other to screen valuable features adapted to different interacted inputs.

Cross-attention Module. We employ self-attention networks as the cross-attention module to explicitly capture the dependencies between any two words and learn the inner structure information of sequences to ensure the deep interaction between the two sequences, which could be formalized as:

$$\mathbf{H} = \mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \quad (1)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are query, keys, and value matrices, respectively. In our setting, $\mathbf{Q} = e^r_{all}$ and $\mathbf{K} = \mathbf{V} = e_c$, d_k is the size of hidden units of BiLSTM, which equals to $2h$.

To enhance the parallelism of the networks, multi-head attention linearly projects the queries, keys, and values m times by using different linear projections, and then performs the scaled dot-product attention in parallel. Finally, the processed results are concatenated and once again projected to obtain the new representation. Formally, the multi-head attention could be expressed as:

$$\begin{aligned} \mathrm{head}_i &= \mathbf{Attention}(\mathbf{Q}\mathbf{W}^Q_i, \mathbf{K}\mathbf{W}^K_i, \mathbf{V}\mathbf{W}^V_i) && (2)\\ \mathbf{H}_s &= \mathbf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \mathbf{Concat}(\mathrm{head}_1, \mathrm{head}_2, ..., \mathrm{head}_m)\mathbf{W}^o && (3) \end{aligned}$$

where $\mathbf{W}^Q_i \in \mathbb{R}^{2h\times d_1}$, $\mathbf{W}^K_i \in \mathbb{R}^{2h\times d_1}$, $\mathbf{W}^V_i \in \mathbb{R}^{2h\times d_1}$, and $\mathbf{W}^o \in \mathbb{R}^{2h\times 2h}$ are all trainable parameters and d_1 is $2h/m$. $\mathbf{H}_s \in \mathbb{R}^{k\times 2h}$ is the output vectors of cross-attention module, i.e., the interactive features between the claim and relevant articles.

Gate Affine Absorption Module. Considering that the interactive features belong to the shared features between claims and relevant articles, which lack the context of their respective sequences, we develop gate affine absorption module to enable the model to capture context-aware interactive features for both sequences, so as to adaptively focus on the prominent semantic segments that are widely concerned by users in claims, and the semantic features that provide users' holistic opinions in relevant articles, respectively.

Specifically, take the gate G1 (as shown in Figure 2) aiming at the claim for example, we first employ activation function to enhance the nonlinear features of the encoding of the claim e_c and the shared interactive features $\mathbf{H}_s$, and then apply linear transformation to map e_c and $\mathbf{H}_s$ to obtain the mapping context-aware claim vector $\alpha(e_c)$, scaling vector $\beta(\mathbf{H}_s)$, and shifting vector $\gamma(\mathbf{H}_s)$, respectively. Next, 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 the key semantics adapted to the claim. Finally, the shifting vector adjusts slightly to the shared features to achieve the appropriate fusion. Formally, the process can be formalized as follows:

$$\begin{aligned} \mathrm{t}_c &= \mathbf{tanh}(\mathbf{W}_c e_c + \mathbf{b}_c) \\ \mathrm{t}_s &= \mathbf{tanh}(\mathbf{W}_s \mathbf{H}_s + \mathbf{b}_s) \\ \alpha(e_c) &= \mathbf{W}_\alpha t_c + \mathbf{b}_\alpha \\ \beta(\mathbf{H}_s) &= \mathbf{W}_\beta t_s + \mathbf{b}_\beta \\ \gamma(\mathbf{H}_s) &= \mathbf{W}_\gamma t_s + \mathbf{b}_\gamma \\ \mathbf{H}^c_s &= f(e_c, \mathbf{H}_s) = \sigma(\beta(\mathbf{H}_s)) \odot \alpha(e_c) + \gamma(\mathbf{H}_s) \quad (4) \end{aligned}$$

where all $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters. $\odot$ denotes element-wise multiplication. Particularly, the gate G2 aiming at all relevant articles is the same as the gate G1, and $\mathbf{H}_s^r$ is the outputs of the gate G2, i.e., the users' holistic opinions in all relevant articles.

3.3 Fine-grained Conflict Discovery layer

In order to discover semantic conflicts aiming at each relevant article, we construct conflict discovery layer consisting of cross-attention module to achieve full interaction and fusion between the users' holistic opinions $\mathbf{H}_s^r$ and local single relevant article e_i^r, where the cross-attention module has been introduced in Section 3.2:

$$\mathbf{H}_i^{clf} = \textbf{Attention}(\mathbf{H}_s^r, e_i^r, e_i^r) \tag{5}$$

where $\mathbf{H}_i^{clf}$ refers to the conflict semantics captured from the i-th relevant article.

3.4 Evidence-aware Coherence Layer

To eliminate the noise irrelevant to the claim in the conflicts of all relevant articles, and infer which conflicts discuss the questioned core fragments of claims, we build evidence-aware coherence layer to measure the coherence between the core semantic segments of the claim and the potential conflicts of each article.

Thus, we first employ two BiLSTMs to respectively encode the core segments $\mathbf{H}_s^c$ of the claim and the conflict $\mathbf{H}_i^{clf}$ of the i-th relevant article, and represent the final hidden states of both BiLSTMs as m_s^c and m_i^{clf}, respectively.

Then, we rely on attention mechanism to match the salient features between the claim and the i-th relevant article. For the j-th word in the i-th article, we have:

$$u_{ij}^{cc} = \mathbf{H}_s^c \, \mathbf{H}_{ij}^{clf} \tag{6}$$

$$\beta_{ij}^{cc} = \textbf{exp}(u_{ij}^{cc}) / \sum_{l=1}^{L_i} \textbf{exp}(u_{il}^{cc}) \tag{7}$$

where $\mathbf{H}_{ij}^{clf}$ is the j-th word in the conflict $\mathbf{H}_i^{clf}$ of the i-th article, u_{ij}^{cc} and β_{ij}^{cc} could be regarded as the raw and normalized association measure of the j-th word in the conflict of the i-th article to the whole claim sequence. Therefore, we obtain a claim-guided conflict representation v_i^{clf} as:

$$v_i^{clf} = \sum_{l=1}^{L_i} \beta_{ij}^{cc} \, \mathbf{H}_{il}^{clf} \tag{8}$$

In order to strengthen the conflict semantics closely related to core semantic fragments of the claim, we conduct an element-wise summation to combine the claim-guided conflict v_i^{clf} and encoding conflict m_i^{clf}:

$$n_i^{clf} = v_i^{clf} \oplus m_i^{clf} \tag{9}$$

where n_i^{clf} is the new composite representation for the conflict of the i-th article, $\oplus$ is the element-wise summation.

Next, we concatenate n_i^{clf} and m_s^c and pass them into a fully-connected layer to get a low-dimensional prediction vector for coherence representation between the core segments of the claim and conflicts of relevant articles:

$$s_i^{cc} = \textbf{MLP}([n_i^{clf}; m_s^c]) \tag{10}$$

For all relevant articles, we conduct similar operations introduced above to obtain the coherence prediction vectors, i.e., s_1^{cc}, s_2^{cc}, ..., s_R^{cc}, respectively.

Finally, we adopt concatenation operation to fuse them and predict the probability distribution by the following equation:

$$p = \textbf{softmax}(\mathbf{W}_p[s_1^{cc}; s_2^{cc}; ..., s_R^{cc}] + \mathbf{b}_p) \tag{11}$$

We train the model to minimize cross-entropy error for a training sample with ground-truth label y:

$$loss = -\sum y \textbf{log} p \tag{12}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We adopt two widely-used competitive datasets released by Popat et al. (2018) for evaluation. Their details are shown as follows:

Snopes Dataset. Snopes[2] possesses 4,341 claims and corresponding 29,242 relevant articles that include opinions on claims retrieved from 3,267 domains by Bing search API. Each claim in Snopes is labeled as true and false.

PolitiFact Dataset. PolitiFact[3] has 3,568 claims and 29,556 relevant articles associated with 3,028 domains. Each claim belongs to one of six credibility ratings in PolitiFact: true, mostly-true, half-true, mostly-false, false, and pants-on-fire. Following Ma et al. (2019), we incorporate mostly true, half true, and mostly false into mixed, and merge false and pants on fire as false.

Evaluation Metrics. We utilize micro-/macro-averaged F1, class-specific precision, recall, and F1-score as evaluation metrics. We hold out 10% of the claims in the two datasets as development set for tuning the hyper-parameters, and conduct 5-fold cross-validation on the rest of the claims.

4.2 Settings

We tune all hyper-parameters via a small grid search for the best performance. For preprocessing, we tokenize the sequences and lowercase the tokens. For parameter configurations, the pre-trained BERT-base model is used to initialize word embeddings. The size of embeddings is set as 768. The number R of relevant articles varies with different claims. The length k of claim sequence and that of each relevant article are set to 30, and 120, respectively, while the length of the integrated sequence of all relevant articles varies with the number of relevant articles. In self-attention networks, attention heads and blocks are set to 6 and 4, respectively, and the dropout of multi-head attention is set to 0.5. Additionally, the initial learning rate is set to 0.001. We use L2-regularizers with the fully connected layers as well as dropout and set it to 0.6, and the mini-batch size is 64.

[2]collected from http://www.snopes.com

[3]collected from https://www.politifact.com

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
AIFN	0.812	0.767	0.614	0.673	0.642	0.877	0.863	0.870	0.527	0.493	0.499	0.631	0.347
Ours	**0.842**	**0.791**	0.620	**0.795**	**0.697**	**0.897**	**0.898**	**0.897**	**0.561**	**0.515**	**0.520**	**0.656**	**0.365**

Table 1: Performance comparison of EVIN 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
-shared	0.813	0.765	0.587	0.758	0.662	0.855	0.858	0.856	0.534	0.490	0.493	0.630	0.334
-gate	0.825	0.783	0.608	0.782	0.684	0.879	0.883	0.881	0.550	0.505	0.510	0.649	0.352
-G1	0.837	0.787	0.616	0.789	0.692	0.892	0.893	0.892	0.554	0.511	0.515	0.652	0.359
-G2	0.839	0.788	0.617	0.792	0.694	0.894	0.895	0.894	0.557	0.513	0.517	0.653	0.363
-conflict	0.821	0.778	0.597	0.776	0.675	0.868	0.872	0.870	0.546	0.502	0.502	0.641	0.347
-coherence	0.801	0.755	0.574	0.749	0.650	0.847	0.846	0.846	0.527	0.483	0.481	0.621	0.326
EVIN	0.842	0.791	0.620	0.795	0.697	0.897	0.898	0.897	0.561	0.515	0.520	0.656	0.365

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

4.3 Performance Evaluation

Comparative baselines We compare the proposed method with some state-of-the-art baselines, including:

- SVM (Ma et al. 2019): A linear SVM based on various handcrafted features for credibility evaluation.
- CNN (Wang 2017): A CNN-based model learning n-grams features from word sequences relying on different window sizes for fake news detection.
- LSTM (Rashkin et al. 2017): The LSTM model capturing hidden representation from word sequences for detection.
- DeClarE (Popat et al. 2018): The interactive model for debunking claims capturing word-level semantic conflicts as interpretable evidence.
- HAN (Ma et al. 2019): Hierarchical attention networks relying on coherence learning to obtain sentence-level semantic conflicts for interpretable claim verification.
- AIFN (Wu and Rao 2020): Adaptive interaction fusion networks discovering multi-feature conflicts from the perspective of semantics and emotions between news and comments.

We implement our models with Pytorch[4], HAN and DeClarE with Theano[5]. We use the original codes of the other baselines. And in CNN-based and LSTM-based models, we only adopt claim content without considering external resources.

[4]https://pytorch.org

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

Overall Performance The results when performing credibility classification on the two datasets are shown in Table 1.

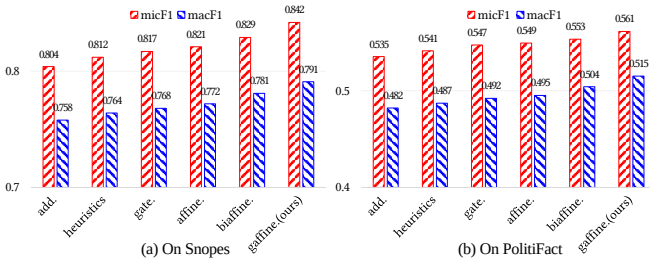
First, among the baseline algorithms, we observe that the deep learning methods (i.e., CNN and LSTM) perform superior than these using hand-crafted features (i.e., SVM). It is not surprising, since the deep learning models can learn high-level hidden representations of claim content. This demonstrates the importance and necessity of relying on neural networks to learn claim content for verification.

Second, EVIN outperforms the traditional neural networks (like CNN and LSTM) in terms of all measures. The reason is that traditional neural networks only focus on claim content to improve performance, but lack exploring the semantic conflicts between claims and relevant articles. This confirms the effectiveness of capturing semantic conflicts for credibility evaluation.

Finally, EVIN achieves more remarkable performance than DeClarE, HAN, and AIFN, showing at least 8.6%, 3.8%, and 3.4% boost in micF1 on the two datasets, respectively. Since the latter three baselines concentrate on securing semantic conflicts with different granularity, such as word level, sentence level, and multi-feature level, as explanations of claim verification, our EVIN puts emphasis on purifying and inferring valid and available evidence from semantic conflicts to interpret verification results. This proves the effectiveness of our model in capturing the conflicts that question the core semantics of claims as evidence.

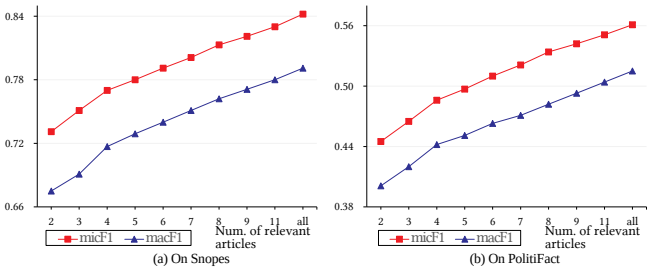
4.4 Discussion

Ablation Study To investigate the effect of each component in EVIN, we conduct ablation analysis by removing different



(a) On Snopes (b) On PolitiFact

Figure 3: Comparison of the adaptive ability of our gate affine transformation with several existing adaptive strategies.



(a) On Snopes (b) On PolitiFact

Figure 4: Comparison of EVIN under different claims with different number of relevant articles.

(a) Evid.(false) (b) Evid.(true) (c) Orign.(false) (d) Orign.(true)

Figure 5: The most frequent words based on the captured evidence and the original data on Snopes. (a) The evidence against false claims, (b) The evidence for true claims, (c) The original false claims, and (d) The original true claims.

modules of our model. We employ -shared, -gate, -G1, G2, -conflict, and -coherence to respectively denote the removal of the following components: co-interactive layer, gate affine absorption, the gate G1, the gate G2, fine-grained conflict discovery layer, and evidence-aware coherence layer. Table 2 summarizes the empirical results on Snopes and PolitiFact. The results show that: first, the removal of different components suffers different degrees of such performance degradation, which reflects the effectiveness of each component. Second, the performance of the model without evidence-aware coherence layer is subjected to a large reduction, indicating that considering the coherence between core semantic segments of claims and semantic conflicts of relevant articles to infer available evidence contributes to boosting the final performance. Finally, as a part of co-interactive shared layer, gate affine absorption module shows at most 1.7% performance degradation in micF1 on the two datasets, which fully embodies the key role of gate affine absorption.

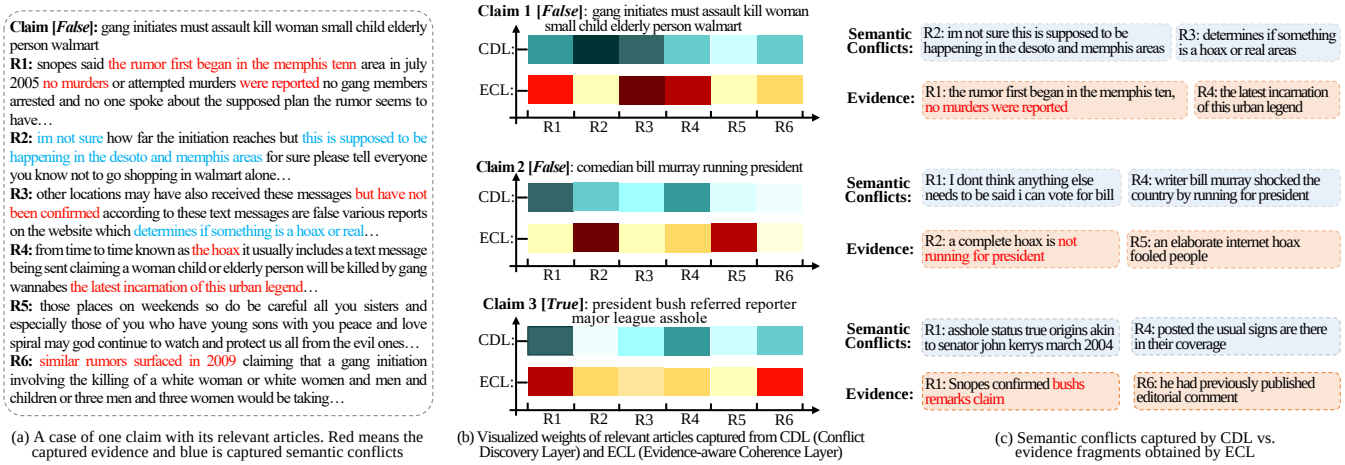
Evaluation of Gate Affine Absorption We know that co-interactive shared layer can adaptively focus on the crucial semantics of claims and relevant articles respectively with the help of gate affine absorption (a.k.a. gaffine.). In order to further evaluate the adaptive advantages of gaffine., we compare it with the following adaptive strategies: **add.** means additions. **heuristics** is matching heuristics (Mou et al. 2016). **gate.** is attentional feature-based gate (Margatina, Baziotis, and Potamianos 2019). **affine.** denotes attentional affine transformation (Margatina, Baziotis, and Potamianos 2019). **biaffine.** is biaffine attention (Ma et al. 2019). These baselines replace our gate affine absorption to combine the shared interactive features with the encoding of claims or that of relevant articles, respectively. The experimental results are presented in Figure 3. We can find that: first, gate. relying on filtering irrelevant features gains better performance than add. and heuristics. And biaffine. capturing valid invariance-based features by aid of two affine transformations shows the best performance in all baselines. Second, our model with gate affine absorption performs more excellent performance than other adaptive models and the improvements are between 0.8% and 3.8% in micF1 on the two datasets. These results demonstrate the ability of our gate affine absorption to capture core semantics adaptively.

Effect of the Number of Relevant Articles Figure 4 provides the performance of EVIN under different claims with different number of relevant articles. We learn that as the number of relevant articles increases, the performance of EVIN improves, and when the claims with less than three relevant articles, the model achieves the most unsatisfactory performance, only showing 75.1% and 46.5% performance in micF1 (at least 9.1% degradation than the model on all articles) on the two datasets, respectively. The reason might be that the fewer relevant articles are, i.e., the fewer users participate in the discussion of the claims, which is easier to weaken the exposure of the false parts of claims, making it difficult for relevant articles to detect these false parts, which correspondingly leads to fewer semantic conflicts in the relevant articles. To some extent, this shows that our model does not perform well in the relevant articles with less conflicts. This also implies that our model is not suitable for early claim verification.

4.5 Interpretability Analysis

In order to evaluate the quality of evidence captured by EVIN and make the verification results easy-to-interpret by this evidence, we illustrate intuitively from the following three perspectives.

Interpretability on Word-level Distribution of Evidence We first visualize the most frequent words of evidence captured by EVIN with word cloud to compare the word distribution between the false claims and true claims on Snopes, as exhibited in Figure 5. We observe that captured evidence against false claims strengthens more skeptical or refuting words, such as 'rumor', 'suspected', and 'uncovered' (Figure 5 (a)), and the evidence for true claims contains more supportive and objective words, like 'confirmed', 'true', and 'reason' (Figure 5 (b)). Nevertheless, there are no similar patterns between the word distributions of true claims and that of false claims aiming at all relevant articles (Figure 5 (c) and (d)). On the whole, the captured evidence learns a wealth of credibility-indicative features, which reveals its superior quality.

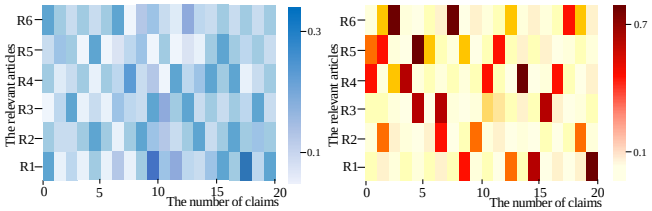


(a) A case of one claim with its relevant articles. Red means the captured evidence and blue is captured semantic conflicts

(b) Visualized weights of relevant articles captured from CDL (Conflict Discovery Layer) and ECL (Evidence-aware Coherence Layer)

(c) Semantic conflicts captured by CDL vs. evidence fragments obtained by ECL

Figure 6: Visualized results of captured evidence (by evidence-aware coherence layer) and semantic conflicts (by conflict discovery layer) based on three cases of claims with their relevant articles.



(a) Weight visualization of relevant articles of 20 claims labeled as True

(b) Weight visualization of relevant articles of 20 claims labeled as False

Figure 7: The weight visualization of relevant articles of 20 claims respectively labeled True or False.

Interpretability on Fragment-level Semantics of Evidence To intuitively compare which of the two, i.e., the conflicts that question the core semantics of claims (the captured evidence) and general semantic conflicts, is more effective in explaining the verification results, we visualize the outputs of both conflict discovery layer and evidence-aware coherence layer at fragment level. Specifically, we first look up these elements with the largest values from the entire outputs of the two layers, and then these elements are mapped into the corresponding values in input embeddings so that we could find the specific tokens. The visualization results are depicted in Figure 6. From Figure 6(a) and (c), we learn that the obtained semantic conflicts contain broader semantics, including ambiguous argument, such as 'I'm not sure' and 'if something is a hoax or real', while the captured evidence not only focuses on the core semantics of the claim, like 'no murders' in claim 1 (red), but also provides useful explanations precisely, such as 'the latest incidence of this urban legend'. From Figure 6(b) and (c), we grasp that the captured evidence fragments are capable of supporting the credibility of claims. For instance, in claim 2, the evidence fragment 'an elaborate internet hoax fooled people' refutes the claim 'comedian bill murray running president' as false. These prove that the captured evidence effectively reveals the false parts of the claim, which is superior to the semantic conflicts in interpreting the credibility results.

Interpretability on Sentence-level Articles of Evidence We map the attention weights of EVIN to different relevant articles (i.e., sentence-level articles) for false and true claims, as shown in Figure 7, which specifically visualizes the attention maps for 20 sampled claims with their relevant articles on Snopes. We can find that the distribution of attention weights for the relevant articles of true claims is more uniform, whereas that of false claims is sparser, in which certain sentences are highly focused. These sentences contain rich questioning voices (according to the last subsection), which conveys that the relevant articles rich in controversial features can be effectively concerned by EVIN. Interestingly enough several articles of false claims receive little attention, showing weak correlation between these articles and their claims, which indicates that the relevant articles of false claims involve more noise features (like, commercial ad., spams, etc.) than these of true claims. This may well provide salient credibility-indicative features for understanding the differences between true and false claims.

5 Conclusion

In this paper, we proposed to investigate the important problem of interpretable claim verification. We endeavored to infer and distill the conflicts that questioned the core semantics of claims and served as the available evidence to explain the verification results. Our model first focused on the core semantic fragments of claims and the users' opinions on claims in the relevant articles, then finely-grained captured semantic conflicts in the opinions for each relevant article, and finally inferred the consistency between these conflicts and core semantics of claims to grasp available evidence. By the evidence, our model thus showed significantly improved performance in claim verification on two competitive datasets and made the verification results easy-to-interpret. In the future, we plan to explore the changes of semantic conflicts under dynamic features by taking into temporal attributes of relevant articles or comments consideration to boost the interpretability of verification. Besides, we try to introduce few-shot learning to address the problem of early claim verification.

6 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, O.; Bhowmik, D.; and Zargari, S. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP*.

Allen, J.; Howland, B.; Mobius, M.; Rothschild, D.; and Watts, D. J. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* 6(14): eaay3539.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, 675–684.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363(6425): 374–378.

Gröndahl, T.; and Asokan, N. 2019. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *ACM Computing Surveys (CSUR)* 52(3): 45.

Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*.

Karimi, H.; and Tang, J. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. 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)*, 3432–3442.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The Web Conference*, 2915–2921. ACM.

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ma, J.; Gao, W.; Joty, S.; and Wong, K.-F. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *ACL*, 2561–2571.

Ma, J.; Gao, W.; Joty, S.; and Wong, K.-F. 2020. An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11(4): 1–28.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the The Web Conference 2018*, 585–593.

Margatina, K.; Baziotis, C.; and Potamianos, A. 2019. Attention-based Conditioning Methods for External Knowledge Integration. In *ACL*.

Mou, L.; Men, R.; Li, G.; Xu, Y.; Zhang, L.; Yan, R.; and Jin, Z. 2016. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *ACL*, 130–136.

Pennycook, G.; and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116(7): 2521–2526.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 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*, 22–32.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2931–2937.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 312–320.

Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; and Liu, H. 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*, 436–439.

Silverman, C. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed news* 16.

Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426.

Wei, P.; Xu, N.; and Mao, W. 2019. Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity. 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)*, 4789–4800.

Wu, L.; and Rao, Y. 2020. Adaptive Interaction Fusion Networks for Fake News Detection. In *ECAI*.

Zhou, X.; Jain, A.; Phoha, V. V.; and Zafarani, R. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice* 1(2): 1–25.

Zhou, X.; and Zafarani, R. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter* 21(2): 48–60.