ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240649

Search, Examine and Early-Termination: Fake News Detection with Annotation-Free Evidences

Yuzhou Yang[a], Yangming Zhou[a], Qichao Ying[a,b], Zhenxing Qian[a,*] and Xinpeng Zhang[a,c]

[a]Fudan University
[b]NVIDIA
[c]Shanghai University

Abstract. Pioneer researches recognize evidences as crucial elements in fake news detection apart from patterns. Existing evidence-aware methods either require laborious pre-processing procedures to assure relevant and high-quality evidence data, or incorporate the entire spectrum of available evidences in all news cases, regardless of the quality and quantity of the retrieved data. In this paper, we propose an approach named **SEE** that retrieves useful information from web-searched annotation-free evidences with an early-termination mechanism. The proposed SEE is constructed by three main phases: **S**earching online materials using the news as a query and directly using their titles as evidences without any annotating or filtering procedure, sequentially **E**xamining the news alongside with each piece of evidence via attention mechanisms to produce new hidden states with retrieved information, and allowing **E**arly-termination within the examining loop by assessing whether there is adequate confidence for producing a correct prediction. We have conducted extensive experiments on datasets with unprocessed evidences, i.e., Weibo21, GossipCop, and pre-processed evidences, namely Snopes and PolitiFact. The experimental results demonstrate that the proposed method outperforms state-of-the-art approaches.

1 Introduction

The ubiquitous availability and accessibility of the Internet have reshaped the way people obtain and engage with information. Yet, it has concurrently paved the way for the rapid propagation of fake news, which can swiftly amass momentum on social media and various online platforms. The propagation of false information has the potential to yield ill-informed perspectives, with serious implications including public opinion manipulation, concealing the truth, and the incitement of crimes.

Fake News Detection (FND) involves analyzing the probability of news containing misconducting information. Early methods mainly utilize low-level coarse statistical analyses of news content to estimate the veracity of news, e.g., punctuation, lexical statistics [5] or grammar [8]. With the evolution of machine learning techniques, hand-crafted pattern analyzers have been largely supplanted by deep networks. The prevailing methodology is to prepare thousands of real and fake news samples, extracting features from each sample, and constructing a classifier to establish a binary classification boundary. The literature has had success stories from either mining linguistic features, e.g., pragmatic pattern [6], writing style [34], sentiment [36], or visual features, e.g., image quality [11], forgery artifacts [4], joint spatial-temporal features [21]. Nevertheless, these methods still rely exclusively on pattern analysis of static news, where the learned abnormal patterns may merely capture the traits of news forgeries within specific datasets or limited temporal windows, neglecting the evolution of attributes of fake news. Besides, attackers could circumvent the pattern analyzers by mimicking the style of real news.

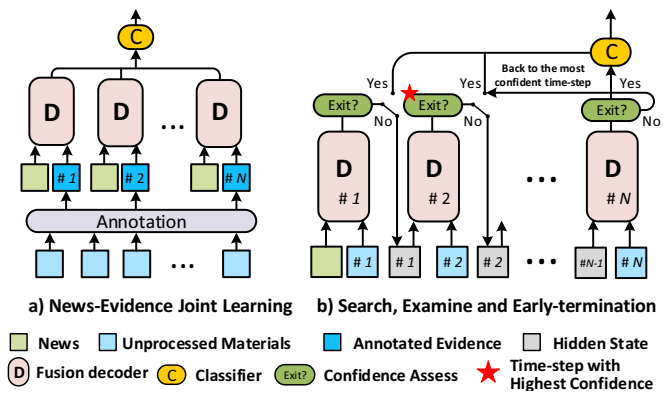

Figure 1. Comparison of methodology between a) previous news-evidence joint learning schemes, typically requiring human-participated evidences and utilizing them altogether, and b) the proposed "Search, Examine and Early-termination" (SEE) scheme, which is capable of utilizing web-searched materials as evidences without annonating and sequentially examining the news alongside each piece of evidence with Early-termination. # denotes the serial number.

Recent works [24, 28, 35, 16, 10] have prominently recognized *evidence* as a premier element in fake news detection apart from patterns. This recognition stems from the observation that individuals tend to search for related news as references during the decision-making process. Researchers have developed many evidence-based FND schemes, where the underlying methodology can be categorized into two types. The first is *high-quality dataset contribution*, which provides cleaned news-evidence pairs and the label of the news is determined by whether the majority of conclusions from the evidences match that of the news. Famous examples of such datasets include PolitiFact [22], Snopes [19], etc. The second is *tailored network design*, which explores enhanced strategies for the news-evidence joint learning [16, 29, 30, 31]. These methods usually refine and fuse the representations of all evidence articles together with that

* Corresponding Author. Email: zxqian@fudan.edu.cn.

of the news [20, 29], somehow resembling a boundary from the joint distribution of the news and evidences based on the aforementioned datasets.

Despite the efforts made in evidence-aware FND, these methods either require laborious pre-processing operations to secure relevant and high-quality evidences during both the training and inference stages, or often utilize all evidences as a supplement to the news, regardless of their quality and quantity. In real-world applications, the Internet is capable of providing an exploding amount of similar yet unfiltered materials as evidences for a given news. It presents challenges in calculating similarity or manually scrutinizing each piece of evidence for quality assurance. Furthermore, is it beneficial to examine as many pieces of evidences as possible? Not always. When individuals observe the retrieved queue of evidences, a tendency is to scrutinize it sequentially and decisively: first focusing on the content of the evidences starting from the leading ones, next relating each to the news with comparing and reasoning, and finally quitting reading further when they are confident to make a decision. It motivates us to study ways of more efficient utilization of evidences that devoid of annotating, and ensure resilience to low-quality or less-related retrieved evidences.

We propose a new fake news detection approach with annotation-free evidences and early-termination. Our method, **SEE**, mainly innovates the inference procedure, which includes three main steps. The first is **Search**, where we search for online materials using the news as query and directly using their titles as evidences without any annotating or filtering procedure. The second is **Examine**, where we encode news and evidences into representations, and employ separate transformer-based *decoders* [27] with joint self- and cross-attention for sequential news-evidence information fusion. The third is **Early-Termination**, where we equip the model with a shared *confidence assessor* that in each time-step, i.e., index of the evidence in the queue, determines either to continue examining more evidences or to provide a prediction with adequate confidence. Fig. 1 depicts the mechanism of our method, in comparison with the previous paradigm of direct joint news-evidence learning. We design a two-stage training mechanism, where we first train the decoders (or feature extractors) and the ultimate binary classifier for useful representation mining from each time-step. Then we fix these networks to train the confidence assessors where the target is one minus the distance between the predicted result in each time-step and the label of the news.

To verify the proposed method, we collect evidences via the Microsoft Bing API [17] for the news in two famous FND datasets, namely, Weibo21 [18] and GossipCop [25], and refrain from making any additional quality control to the retrieved evidences. We apply many state-of-the-art methods on the datasets with the same evidences and experiments show that our method provides the highest average accuracy. Besides, we verify the robustness of our SEE approach by alternation, shuffling, removal, or void insertion of evidences, which show little impact on the overall FND performance. Moreover, we conduct experiments on two more datasets already with processed evidences, namely, Snopes [19] and PolitiFact [20], which also show leading results.

The contribution of this paper is three-folded:

- We propose a fake news detection approach that is capable of retrieving useful information from annotation-free online retrieved materials, which saves laborious annotating or other pre-processing procedures for quality control over utilized evidences.
- We sequentially feed the evidences and devise an early-termination mechanism to use the evidences more efficiently. The assessor gives a confidence score in each time-step and determines when to quit reading further evidences and move on to giving a prediction.
- Experiments on datasets without pre-processed evidences, i.e., Weibo21 and GossipCop, or with pre-processed evidences, i.e., PolitiFact, Snopes, consistently demonstrate the effectiveness of the proposed SEE compared to the state-of-the-art methods.

2 Related Works

2.1 Pattern-based Fake News Detection

Numerous studies have been conducted to develop automatic methods that detect fake news without considering external evidence. BiGRU [15] and TextCNN [37] respectively use a bidirectional GRU and a 1-D CNN module for feature extraction from the text. BERT [12] is also frequently utilized as FND baselines where the parameters are kept tunable and the classifier works on the CLS token. Ajao et al. [2] propose to analyze the sentimental characteristics of fake news, benefiting some latter works [36]. M^3FEND [40] uses a memory bank to enrich domain information of the news to assist with detection. Also, there are a list of multimodal FND methods that further consider the joint distribution of image and text. Chen et al. [7] use VAEs to compress the images and texts and learn to minimize the Kullback-Leibler (KL) divergence for correctly matched image-text pairs contrastively. Ying et al. [33] extract features from multiple views and design a scoring mechanism to adaptively adjust the weight of each view in the final decision. Zhou et al. [38, 39] design multi-modal fusion mechanisms with pre-trained models. Nevertheless, these methods mainly rely on pattern analysis of static news, neglecting the possible evolvement of characteristics of fake news.

2.2 Evidence-aware Fake News Detection

Many high-quality fake news datasets with evidences are proposed. Snopes [19] and PolitiFact [22] contains retrieved evidence articles for each claim by issuing each claim as a query to the Microsoft Bing API, articles are processed by filtering out those related to Snopes and PolitiFact websites and calculating relevance scores to decide on their usage. Similar datasets are FEVER [26], Emergent [9], etc. In contrast, other famous datasets such as Weibo21 and GossipCop do not provide evidences and therefore have only been applied for pattern-based, rather than evidence-aware, FND. Besides, many tailored detection networks are proposed on top of these datasets. DeClarE [20] averages signals from external evidence articles and concatenates them with the language of the articles and the trustworthiness of the sources. Vo et al. [29] proposes to use hierarchical multi-head attention network to combine word attention and evidence attention. CCN [1] leverages both the image caption and text for online evidences, and detects the consistency of the claim-evidence (text-text and image-image), in addition to the image-caption pairing. Xu et al. [32] introduces GET that applies a Graph Neural Network (GNN) to capture long-distance semantic dependency among the news and evidence articles. However, previous methods either require laborious pre-processing operations towards the evidences, or utilize all evidences for news cases regardless of the quality and quantity. We propose a new FND approach with annotation-free evidences and early-termination.

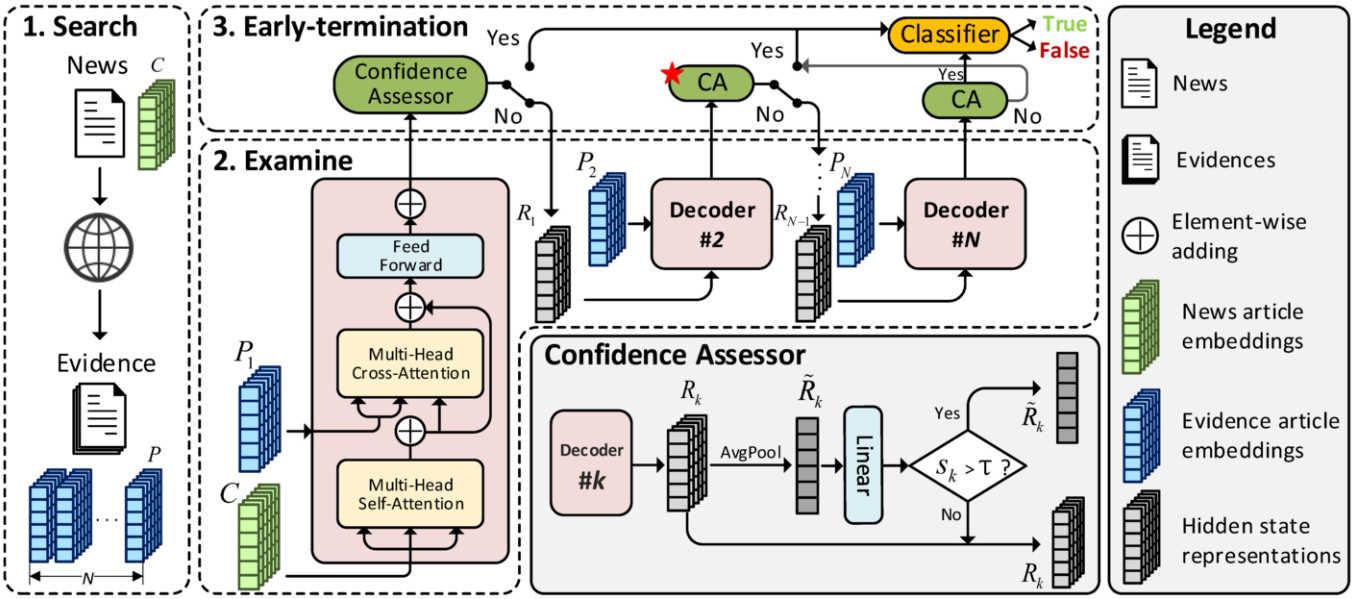

Figure 2. Network design of SEE, our fack-checking FND scheme with annotation-free evidences. SEE includes three main stages: **1. Search**: online materials are retrieved as evidences without any annotating or filtering. **2. Examine**: the decoders sequentially examine the news alongside each piece of evidence via self- and cross-attention to acquire more comprehensive information. **3. Early-termination**: a shared confidence assessor reduces the hidden states in each time-step into confidence scores and determine either to continue examining more evidences or to terminate and predict.

3 Proposed Approach

Fig. 2 depicts the pipeline of the proposed SEE approach. It consists of three stages, namely, 1) *S*earching online materials using the news as query and directly using their titles as the retrieved evidences without any annotating or filtering procedure, 2) sequentially *E*xamining the news alongside with each piece of evidence via attention mechanisms, which produce new hidden states with richer information, and 3) using *E*arly-termination within the examination loop by assessing whether sufficient confidence for a correct FND prediction exists.

3.1 Search: Evidence Preparation

Let the input news be $\mathcal{N} = [\mathbf{c}, \mathcal{E}] \in \mathcal{D}$, where $\mathbf{c}, \mathcal{E}, \mathcal{D}$ are the text of the news, a queue of the corresponding retrieved evidences and the news dataset, respectively. $\mathcal{E} = [e_1, e_2, ...]$ can be either provided by the dataset along with $\mathbf{c}$, e.g., PolitiFact and Snopes, or prepared by the users via online searching, e.g., Weibo21 and GossipCop. We circumvent laborious annotating or other pre-processing operations on $\mathcal{E}$ for quality control. For evidence retrieval, our preliminary is that users trust a certain credibility control of the applied search engine for collecting evidence through the corresponding API. Here, we apply the popular Microsoft Bing to collect evidences. To benchmark "annotation-free" evidence retrieval, we purposefully exclude any pre-processing operations other than pasting the news content into the search box & clicking "Search" for all datasets. We turn each piece of the retrieved material into evidence via recording its title, in accordance with Snopes and PolitiFact, and store in order at most N evidences in each $\mathcal{E}$, where N is a tunable hyper-parameter. Exampled $\mathcal{N}$ for Weibo21 and GossipCop are provided in the experiments[1]. Next, we adopt the BERT model [12] as the preliminary feature extractor for both news and the evidences. The representations of the news and evidences are respectively denoted as $C = \text{BERT}(\mathbf{c}) \in \mathbb{R}^{L\times d}$, $P_i = \text{BERT}(e_i) \in \mathbb{R}^{L\times d}$, $i \in [1, N]$, and d denotes the feature dimension.

[1] We will open-source the collected online evidences for Weibo21 and GossipCop after the anonymous reviewing process.

3.2 Examine: Attention-based News-evidence Fusion

We prepare a cascade of N independent transformer decoders [12] with joint self- and cross-attention to sequentially examine the news alongside each piece of evidence and produce new hidden states with richer information. Each decoder is responsible for feature fusion for a typical time-step. The first block takes C and the leading evidence P_1 as inputs, and outputs R_1, which we call *hidden state* of the first time-step. The following blocks take the hidden state by the previous block and the initial embedding of the upcoming evidence to iteratively produce a list of hidden states. We use the attention mechanism and feed-forward layer operation in typical transformers [12], denoted with $\text{Attn}(\cdot)$ and $\text{FFN}(\cdot)$. Specifically, the attention operation has:

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

where Q, K, V denotes the query, key, and value matrix.

In the decoders, we combine both self-attention and cross-attention in the examining step for news-evidence interaction. Each decoder consists of three layers: a self-attention block, a cross-attention block, and a feed-forward network. Each layer has a residual connection to the previous layer and is attached to a layer normalization. For simplicity in equations, we use overlines to denote the Layer Normalization operations (LN) [3]. The examination process of a decoder block is:

$$\begin{aligned} L_1 &= \begin{cases} C + \text{Attn}(\overline{C}, \overline{C}, \overline{C}), & \textit{if } k = 1 \\ R_{k-1} + \text{Attn}(\overline{R_{k-1}}, \overline{R_{k-1}}, \overline{R_{k-1}}) & \textit{otherwise} \end{cases}, \\ L_2 &= L_1 + \text{Attn}(\overline{L_1}, \overline{P_k}, \overline{P_k}), \\ R_k &= L_2 + \text{FFN}(\overline{L_2}). \end{aligned} \tag{2}$$

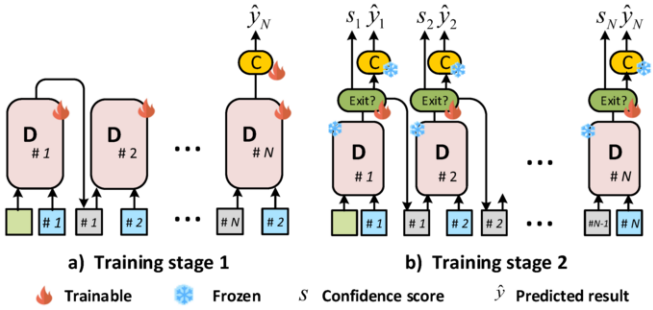

Figure 3. Illustration of the two training stages. In stage 1 we solely emphasize feature extraction from evidences from different indexes in the retrieved queue. In stage 2 the assessor transforms the representations in each time-step into confidence scores for early-exiting on top of fixed feature extractors. The inference stage is depicted in Fig. 1.

Note that each decoder is independent and receives evidence from a specific index in the queue, where the inherent relationship between each index and its statistical relevance of the evidence to the news can be implicitly learned. One can also alternatively employs a shared decoder and prompts it using different indices to save computation. However, this approach leads to decreased accuracy in comparison to using individual decoders, as shown in the ablation studies (Section 4.4).

3.3 *Early-termination: Exit When "Confident"*

Consider that examining all evidences as supplements to the news regardless of the quality and quantity might be sub-optimal, as the existence of low-quality or less-relevant evidences could impact the model performance. We introduce the early-termination step that enables the model to cease further examining more pieces of evidence and proceed to prediction if a confidence threshold is surpassed within each time-step.

In detail, the confidence assessor visits every hidden state $R_k, k \in [1, N]$ and reduces them into confidence scores s_k. We set a hyper-parameter τ as a threshold. τ is set to decide whether to continue the examining step by sending R_k and the next, if exists, evidence into the next decoder, or to terminate and make the prediction, denoted as $\hat{y}$, by sending $\overline{R_k}$ into the classifier. τ is another hyper-parameter, which has a non-negligible effect on both the overall ratio of early-termination and detection performance. We give a detailed analysis of τ in Section 4.2.

During the iterative examining and confidence assessment, we record the hidden state with the highest confidence score, denoted as R'. If the last evidence is reached, we resort to using R' with the highest score to make the prediction $\hat{y}$. The classifier is a three-layer perceptron and a sigmoid function. Therefore, we denote the calculation procedure of $\hat{y}$ as follows.

$$\hat{y} = \begin{cases} \sigma(\mathrm{MLP}(\overline{R_k})), & \textit{if } s_k \geq \tau \\ \sigma(\mathrm{MLP}(\overline{R'})) & \textit{otherwise} \end{cases}. \tag{3}$$

For the confidence assessor, given a hidden state R_k at time-step k, we first reduce the token dimension using average pooling over all tokens and use a fully-connected layer $f(\cdot)$ to reduce it into a logit, which is further mapped within $[0, 1]$ as the confidence score by a sigmoid function $\sigma(\cdot)$.

$$\begin{aligned} \overline{R_k} &= \mathrm{AvgPool}(R_k), \\ s_k &= \sigma(f(\overline{R_k})). \end{aligned} \tag{4}$$

Algorithm 1 Pseudo-codes for the training pipeline.

Input: Training set: $\mathcal{T}$, Validation set: $\mathcal{V}$, Amount of evidences: N, Training epochs: M
Output: Assessor parameters: Ω, Model parameters besides Ω: Θ

for m in range(M): **do**
  Sample $\{\mathbf{c}, \mathcal{E}, y\}$ from $\mathcal{T}$, $\mathcal{E} = [e_1, ..., e_N]$.
  Extract features $C = \mathrm{BERT}(\mathbf{c})$, $P_k = \mathrm{BERT}(e_k)$.
  // Stage One: for Feature Extraction & Detection.
  Let $R_0 = C$.
  for k in range(N): **do**
    $R_k = \mathrm{Decoder}_k(R_{k-1}, P_k)$
  end for
  $\hat{y} = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(R_N)))$
  Compute $\mathcal{L}_{cls}(y, \hat{y})$ and update parameters in Θ.
  // Stage Two: for Early-Exiting.
  for k in range(N): **do**
    $R_k = \mathrm{Decoder}_k(R_{k-1}, P_k)$
    $s_k = \sigma(f(AvgPool(R_k)))$
    $\hat{y}_n = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(R_k)))$
    $y'_k = 1 - |y - \hat{y_k}|$.
  end for
  Compute $\mathcal{L}_{CA}$ and update parameters in Ω.
end for

Algorithm 2 Pseudo-codes for the inference pipeline.

Input: Validation/Test set: $\mathcal{V}$, Amount of evidences: N, Early-exiting threshold: τ
Output: Prediction: y

Sample $\{t, \mathcal{E}\}$ from $\mathcal{V}$, $\mathbf{E} = [e_1, ..., e_N]$.
Extract features $C = \mathrm{BERT}(t)$, $P_k = \mathrm{BERT}(e_k)$.
Let $R_0 = C$, $s_{max} = 0$, $t = 0$.
for k in range(N): **do**
  $R_k = \mathrm{Decoder}_k(R_{k-1}, P_k)$
  $s_k = \sigma(f(\mathrm{AvgPool}(R_k)))$
  if $s_k > \tau$ **then**
    return $y = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(R_k)))$
  else if $s_k > s_{max}$ **then**
    Let $s_{max} = s_k, t = k$.
  end if
end for
return $y = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(R_t)))$

3.4 *Two-staged Training Mechanism*

There is no external label for the confidence assessment process. We propose a two-staged training mechanism where we first solely emphasize on feature extraction from evidences, and then train an assessor to determine when to exit with fixed feature extractors. Soft labels are automatically assigned in each time-step by comparing the current predicted result with the ground-truth FND hard label.

Stage One: Training Feature Extractors and Classifier. The initial phase involves training the feature extractors, i.e. the tunable BERT and decoders, and the classifier. The confidence assessor remains uninvolved during this stage. Specifically, we iteratively feed all N evidences without early-termination, and every decoder generates a unique hidden state in each time-step. The goal is to enable the decoders to extract useful features from the provided news and evidences at different time-steps. The hidden state of the last decoder R_N is then fed to the classifier, which produces $\hat{y}_N$. We update the

networks using the classification loss $\mathcal{L}_{cls}$, where

$$\mathcal{L}_{cls} = y \log(\hat{y}_N) + (1-y) + (1-y)\log(1-\hat{y}_N). \quad (5)$$

The classifier is also jointly trained in the settings where all N evidences are provided. The purpose of this stage is to provide the classifier with adequate information for decision-making as well as encourage the decoders to maximally reserve the most critical information till the last time-step.

Stage Two: Training the Confidence Assessor. After the first stage, we train the confidence assessor to provide confidence scores for each time-step. We fix the models updated in the first stage, i.e., the tunable BERT, the decoders, and the classifier, then only train the confidence assessors where the target is one minus the distance between the predicted result in each time-step and the label of the news. Note that here a shared assessor is employed across all time-steps in that we want the assessor to terminate inference on feeling "confident" without the prior of index. Here, the "confidence" is represented by a predicted score that might not align with human-defined confidence. Specifically, we get the predicted scores $\{\hat{y}_1, ..., \hat{y}_N\}$ in each time-step by respectively sending the corresponding hidden states $\{\mathbf{R}_1, ..., \mathbf{R}_N\}$ into the fixed classifier. The labels of the confidence scores, denoted as y'_k, are defined as one minus the distance between $\hat{y}_k$ and y, as in Eq. 6, i.e., the larger the prediction deviates from the ground truth, the lower the confidence should be, ranging from zero to one.

$$y'_k = 1 - |y - \hat{y}_k|, k \in [1, N]. \quad (6)$$

The confidence assessor is trained to regress y'_k in each time-step, where we use the summed L_1 loss for training:

$$\mathcal{L}_{CA} = \sum_{k=1}^{N} |y'_k - s_k|. \quad (7)$$

Finally, we provide the pseudo-code of detailed training and inference processes of SEE in Algo. 1 and Algo. 2.

4 Experiments

4.1 Experimental Setups

Implementation Details. We use *bert-base-chinese* and *bert-base-uncased* pre-trained models for processing Chinese dataset and English datasets, respectively. The hidden size of word embeddings is 768 ($d = 768$). We unify the length of input news and evidence to a specific length by padding or truncating ($L = 100$). For samples with fewer than N evidence, the missing evidences are treated as blank and filled with [PAD] tokens. To save computation, the representations can be pre-computed, stored on disk and loaded upon training. Our model is trained using a single NVIDIA GeForce RTX 4090 GPU. We use Adam optimizer [13] with default parameters. The batch size is 12. We use a learning rate of 6×10^{-6} for the feature extractors, i.e., fine-tuned BERT and the decoders, and 5×10^{-5} for the rest of the model.

Data Preparation. We use four famous datasets, namely Weibo21 [18], GossipCop [25], Snopes [19], and PolitiFact [20] for our experiments. Weibo21 is a Chinese fake news detection dataset collected from Sina Weibo. News content of GossipCop, Snopes, and PolitiFact are collected from fact-checking websites. Weibo21 and GossipCop do not contain official evidences, so we collect evidence articles for them by the method mentioned in Section 3.1. We conduct a train-val-test split in the ratio of 6:2:2 in accordance with previous methods. Table 1 summarizes these datasets.

Table 1. Statistics of the applied FND datasets. # denotes "the number of".

Dataset	# Total	# True	# False	# Evidence
Weibo21	9,128	4,640	4,488	90,550
GossipCop	22,140	16,817	5,323	218,525
Snopes	4,341	1,164	3,177	29,242
PolitiFact	3,568	1,867	1,701	29,556

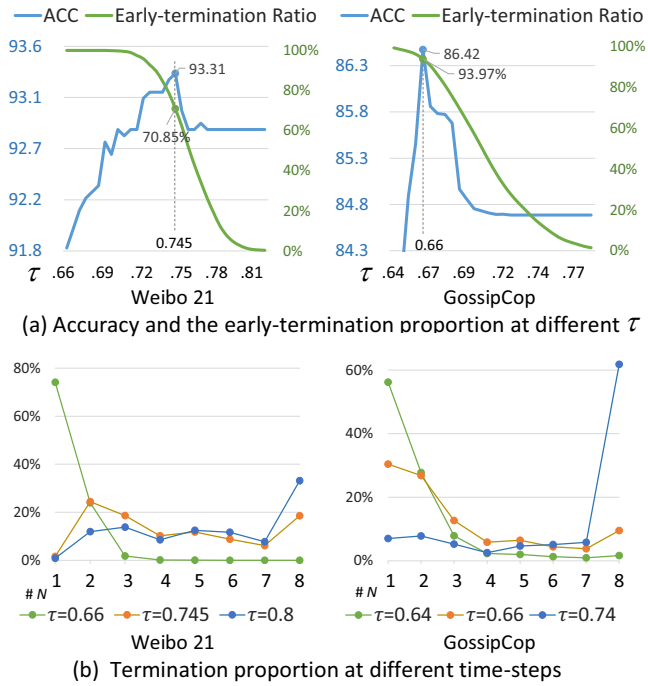

Figure 4. Quantitative analysis of the impact of τ on the performance and early-termination.

4.2 Performance Analysis

Before conducting extensive comparisons with previous state-of-the-arts, we first study the impact of different implementation settings and evidence arrangements. By doing this we aim to elaborate on the selections of key parameters and their impact on the performances, as well as verifying the stability of SEE under different evidence-retrieving situations.

Impact of the Threshold τ on Accuracy and Efficiency. The threshold τ influences early-termination and therefore also the accuracy & efficiency of our method. In the row (a) of Fig. 4, we record the detection accuracy alongside the early-termination rate using varied values of τ. Here the early-termination ratio calculates the probability of over-threshold confidence scores, which is defined as the number of samples that terminate early divided by the total number of samples. For accuracy curves (blue lines), they first rise with the increase of τ, and then descend and stabilize after reaching the peak value. This observation shows that using information from adequate pieces of evidence can be often enough to produce a reliable prediction, which might be counter-intuitive with the point that viewing all of the retrieved evidences is essential and better.

Meanwhile, τ also influences the efficiency of inference, where larger τ indicates that the model would be more cautious during examinations, empirically requiring more evidences on average to produce a determination. We visualize the proportion of termination at each time-step in the row (b) of Fig. 4. The orange curve illustrates the optimal τ on the dataset if we only prevail high accuracy. It gen-

Table 2. Quantitative analysis of the impact of the retrieved results from the view of quality, quantity, and order, detailed description of each setting is provided in section 4.2. We report the accuracy, macro F1 score, and area under ROC (AUC). *: using the pre-filtered evidences within the dataset.

Evidence Adjustments	Weibo21			GossipCop			Snopes*			PolitiFact*		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Most Related Swapped	0.927	0.927	0.927	0.860	0.798	0.783	0.819	0.774	0.766	0.663	0.659	0.662
All Evidences Shuffled	0.926	0.926	0.926	0.856	0.792	0.776	0.798	0.747	0.740	0.667	0.664	0.666
Reverse the Sequence	0.927	0.927	0.927	0.860	0.768	0.782	0.798	0.747	0.739	0.676	0.671	0.674
Most Related Void	0.931	0.931	0.930	0.860	0.796	0.780	0.806	0.748	0.732	0.646	0.632	0.644
Most Related Missing	0.926	0.926	0.925	0.857	0.787	0.765	0.746	0.667	0.657	0.608	0.603	0.607
Limited Evidence (1)	0.905	0.905	0.904	0.852	0.783	0.765	0.770	0.721	0.719	0.676	0.671	0.674
Limited Evidences (3)	0.926	0.926	0.926	0.857	0.788	0.767	0.812	0.758	0.743	0.694	0.691	0.693
Baseline Setting	**0.932**	**0.932**	**0.932**	**0.864**	**0.807**	**0.796**	**0.824**	**0.786**	**0.783**	**0.706**	**0.705**	**0.706**

erally manifests an even distribution, suggesting that the leading evidences only gain moderate extra importance in comparison with the rest of evidences.

Selection of proper τ. We search for the best early-termination threshold τ on the validation set before testing. In the real world, the performance of overall detection accuracy is usually in favored of over inference time, so here we temporarily select τ without additionally considering the trade-off between accuracy and efficiency. According to Fig. 4, $\tau = 0.745$ and $\tau = 0.660$ respectively yield the best accuracy on the validation set of Weibo21 and GossipCop. Besides, the optimal threshold for Snopes is $\tau = 0.715$, and that for PolitiFact $\tau = 0.690$. In practical scenarios, since the four applied datasets are representative and popular, users can select a proper τ ranging from 0.715 to 0.745 for Chinese news, or 0.660 to 0.690 for English news. Moreover, the experiments indicate that even when the threshold is roughly set within an acceptable small range, the caused variation in accuracy is within 2% of the peak value for both datasets.

Impact of the Retrieved Results. Provided with the same query, the searching engine might also retrieval different combinations of materials over time, potentially changing the prediction results of fact-checking based FND methods. Therefore, after the selection of proper τ, we also study the impact of the retrieved results from the view of quality, quantity, and order. The settings are: 1) *Most Related Swapped* presents the leading three (usually most related) materials in the retrieved queue in an arbitrary order, while keeping intact the remaining materials. 2) *All Evidences Shuffled* simulates a more challenging scenario where all materials are in an arbitrary order. 3) *Reverse the Sequence* simulates a scenario in which less relevant evidence is examined first, which is an even more challenging scenario compared to *All Evidences Shuffled*. 3) *Most Related Void* simulates that the first (primary) evidence conveys no valuable information, e.g., due to advertisement or failed retrieval. In this case the first evidence is replaced with a random irrelevant article. 4) *Most Related Missing* simulates the absence of the primary evidence. In this case the first evidence is simply polled out from the queue. 5) *Limited Evidences (n)* simulates that the searching queue only contains a maximum of n evidence. On doing the experiments, we first train our model using the default content and order of evidences in each dataset, and then test them using the above-mentioned settings.

The results are reported in Table 2. Among all adjustments, providing *limited evidences (1 or 3)* decreases performance most evidently, which suggests that considering more evidence is necessary for detection. Adjustments on the order of evidences or the content of the most related evidences result in different performance drop trend on different datasets. For datasets without pre-processed evidences, i.e., Weibo21 and GossipCop, we see that the performance drop is trivial, i.e., usually less than 1%. In contrast, the drop is more noticeable on datasets with pre-processed evidences, i.e., Snopes and PolitiFact, even if only the leading materials are swapped. The finding suggest that the pre-filtering stage might has the ability to indeliberately mislead the model to over-highlight the order of the evidences as well as the role played by the most related evidence. In comparison, using annotation-free evidences suggest a more robust and consistent result less affected by the quality of the first evidence and the order.

4.3 Comparisons

The compared benchmark methods are listed in Table 3. Since our testing datasets might not have been used by some of the baseline methods, we carefully re-implemented them by providing all available required data, e.g., publisher information, propagation graph, etc. In contrast, SEE refrains from using the related additional information other than annotation-free evidences.

Implementation of Baselines. DeClare [20] and GET [32] also require publisher information and other propagation-based information. As such, we do not directly borrow the experimental results from their papers and instead collect related information and carefully retrain the models on the testing datasets. GET involves generating graphs by segmenting words and using pre-trained embeddings, and we implement this on Weibo21 by using jieba [2] and Chinese word vectors pre-trained on weibo [14]. Domain information of news is mandatory for M^{3}FEND, which is not available in GossipCop, Snopes, and PolitiFact. We remove visual modality parts of the CCN [1] and report the results using Sentence-BERT [23] and pre-trained BERT with LSTM versions of it, denoted respectively as CCN-sent and CCN-lstm.

Results. Table 3 presents accuracy, macro F1 score, and area under ROC (AUC) for performance evaluation. SEE achieves the average accuracy on all four datasets, which outperforms the compared benchmark methods. SEE also attains the highest F1 score across all datasets, demonstrating its classification ability. GET exhibits competitive performance with SEE on Snopes and PolitiFact. However, SEE outperforms it notably on datasets without pre-processed evidences, i.e., Weibo21 and GossipCop. Similarly, CCN performs closely to SEE on datasets without pre-processed evidences, yet underperforms on Snopes and PolitiFact. The reason is mainly that GET and CCN are proposed to examine evidence alongside with news at levels of word without information retrieval stage, which demands a high quality of evidence text. As a result, the models are not trained explicitly to utilize un-filtered evidences, which might show different characteristics with the filtered ones. Notably, we see that these evidence-based methods might not significantly outperform methods

[2] https://github.com/fxsjy/jieba

Table 3. Performance comparison between our method with others. *: using the pre-filtered evidences within the dataset. [1]: methods not using evidences, [2]: methods using evidences. −: lacking domain information for experiments.

Method	Weibo21			GossipCop			Snopes*			PolitiFact*		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
BiGRU[1] [15]	0.827	0.827	0.898	0.781	0.783	0.764	0.712	0.615	0.608	0.624	0.623	0.623
TextCNN[1] [37]	0.872	0.872	0.873	0.840	0.757	0.734	0.697	0.502	0.529	0.592	0.592	0.592
BERT[1] [12]	0.918	0.918	0.918	0.855	0.792	0.779	0.733	0.627	0.619	0.604	0.583	0.601
M^{3}FEND[1] [40]	0.922	0.922	**0.975**	0.824	-	-	-	-	-	-	-	-
DeClarE[2] [20]	0.850	0.850	0.850	0.798	0.670	0.667	0.786	0.694	0.677	0.635	0.630	0.631
CCN-sent[2] [1]	0.853	0.852	0.853	0.854	0.783	0.764	0.740	0.658	0.649	0.677	0.676	0.676
CCN-lstm[2] [1]	0.876	0.875	0.875	0.826	0.747	0.734	0.747	0.684	0.679	0.624	0.623	0.625
GET[2] [32]	0.666	0.662	0.667	0.847	0.773	0.754	0.814	0.771	0.767	0.694	0.689	0.692
SEE[2] (Proposed)	**0.932**	**0.932**	0.932	**0.864**	**0.807**	**0.796**	**0.824**	**0.786**	**0.783**	**0.706**	**0.705**	**0.706**

Table 4. Ablation studies of the proposed SEE on network design and training mechanism.

Ablation Settings	Weibo21			GossipCop		
	ACC	F1	AUC	ACC	F1	AUC
Max. allow 6 evidences	0.922	0.922	0.922	0.858	0.788	0.766
Max. allow 10 evidences	0.916	0.916	0.916	0.854	0.779	0.755
Decoders sharing weights	0.883	0.883	0.884	0.847	0.771	0.749
Training models in one go	0.827	0.827	0.898	0.781	0.783	0.764
Concat All Hidden States	0.929	0.928	0.928	0.859	0.798	0.781
Baseline setting (Max. 8)	**0.932**	**0.932**	**0.932**	**0.864**	**0.807**	**0.796**
Ablation Settings	Snopes			PolitiFact		
	ACC	F1	AUC	ACC	F1	AUC
Max. allow 6 evidences	0.821	0.778	0.772	0.704	0.484	0.524
Max. allow 8 evidences	0.751	0.592	0.595	0.636	0.597	0.611
Decoders sharing weights	0.780	0.720	0.710	0.664	0.662	0.665
Training models in one go	0.803	0.748	0.736	0.686	0.681	0.682
Concat All Hidden States	0.818	0.785	0.763	0.701	0.696	0.695
Baseline Setting (Max. 5)	**0.824**	**0.786**	**0.783**	**0.706**	**0.705**	**0.706**

without evidence on Weibo21 and GossipCop, but both show decent performance gain on datasets with pre-processed evidence. Therefore, we could infer that the performance boost of these methods relies partially on the pre-processing stage in these methods. In contrast, the consistently strong performances of SEE on four datasets demonstrate that it overcomes the above-mentioned disadvantages, which could circumvent the time-consuming human-aided filtering.

4.4 Ablation Studies

Enabling Different Amount of Evidences. In Table 4, we vary the maximum quantities of evidences for testing. The performances drop evidently compared to the default setting. It suggests that the training stage is impacted altogether by the quantity of evidences, even though some of which might show less relation with the news. The same performance drop happens to results on Snopes and PolitiFact, suggesting this does not depend on evidence quality. As stage one considers evidences sequentially, and produces detection outcomes solely at the final decoder, an excessive input-to-output span thus damages the training. The input of front decoders may be overshadowed as the data propagates deeper. Therefore, limiting the number of input evidence, which equals limiting the depth of SEE, benefits the performance.

Shared or Independent Decoders. In Table 4, we investigate the utilization of a shared decoder informed by time-step information through positional encodings. The resulting accuracy witnesses declines of 4.9% and 1.7%, 4.4%, and 4.2%, on four datasets respectively, suggesting that assigning a single decoder to capture universal features across evidences sequential locations compromises the examination proficiency of the proposed SEE.

Two-staged or One-go Training. We test the necessity of the two-staged training by altering it with *training the model in one go*, in which we jointly train all components. During training, the hidden state at each time-step is directed into the assessor. In instances where termination is deemed appropriate, the remaining evidences are disregarded, leaving subsequent decoders untrained. The proposed approach suffers from performance drops under this alternative approach on all datasets. According to our observation of details, a substantial number of samples tend to terminate right after the first decoder block, indicating that the confidence assessment holds limited validity. We conclude that jointly training all components damages both the examination and termination abilities of SEE.

Concatenating All Hidden States from Every Time-step. Previous methods mainly utilize concatenation to fuse representations of examined evidences. We exclude the confidence assessment during inference by concating the hidden states at all time-steps, which produces a hidden representation of size $L \times Nd$. This setting underperforms baseline settings, which demonstrates the effectiveness of sequential examination of evidence.

5 Conclusions and Future Works

We introduce SEE, a FND method with Search, Examine and Early-termination based on annotation-free evidences. Our approach incorporates confidence assessment trained on annotation-free evidences for early-termination within examination loops without the effort of human-intervened evidence labelling. Through the method, we are motivated by two key insights, which are then verified by experiments: 1) it is not always necessary to utilize as many evidences as possible to make correct FND prediction, and 2) training models upon well-crafted useful information might mistakenly delude it to highlight the order of all evidences as well as the role played by the most relevant one. Our extensive results underscore the superiority of SEE over state-of-the-art methods, validating its robustness across diverse scenarios. We conclude that SEE excels in distinct evidences utilization and detection capabilities, suggesting that guiding models to assess annotation-free evidences aids in evidence-aware FND.

While we made progress in directly utilizing web-searched raw evidences, the in-depth mechanism of evidence examination as well as that of the early-exiting remains implicit to us. It would be beneficial to further improve the interpretability of FND methods by perhaps utilizing large language models which have zero-shot reasoning abilities for a Q&A.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants U20B2051, 62072114, U20A20178, and U22B2047.

References

[1] S. Abdelnabi, R. Hasan, and M. Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949, 2022.

[2] O. Ajao, D. Bhowmik, and S. Zargari. 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, 2019.

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161, 2020.

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

[6] Y. Chen, N. J. Conroy, and V. L. Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

[7] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905, 2022.

[8] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

[9] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL, 2016.

[10] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, and P. Nakov. Crowdchecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 266–285, 2022.

[11] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2016.

[12] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[13] D. Kingma. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*, 2014.

[14] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-2023.

[15] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

[16] J. Ma, W. Gao, S. Joty, and K.-F. Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics, 2019.

[17] Microsoft. Bing web search api. https://www.microsoft.com/en-us/bing/apis/bing-web-search-api, 2023.

[18] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

[19] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012, 2017.

[20] K. Popat, S. Mukherjee, A. Yates, and G. Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 22–32. ACL, 2018.

[21] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE, 2019.

[22] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 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*, pages 2931–2937, 2017.

[23] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992, 2019.

[24] G. K. Shahi, J. M. Struß, and T. Mandl. Overview of the clef-2021 checkthat! lab: Task 3 on fake news detection. In *CLEF (Working Notes)*, pages 406–423, 2021.

[25] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188, 2020.

[26] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] N. Vo and K. Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 275–284, 2018.

[29] N. Vo and K. Lee. Hierarchical multi-head attentive network for evidence-aware fake news detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 965–975, 2021.

[30] L. Wu, Y. Rao, Y. Lan, L. Sun, and Z. Qi. Unified dual-view cognitive model for interpretable claim verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 59–68, 2021.

[31] L. Wu, Y. Rao, X. Yang, W. Wang, and A. Nazir. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1388–1394, 2021.

[32] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510, 2022.

[33] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, 2023.

[34] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, et al. A convolutional approach for misinformation identification. In *IJCAI*, pages 3901–3907, 2017.

[35] H. Zhang, Q. Fang, S. Qian, and C. Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951, 2019.

[36] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.

[37] Y. Zhang and B. C. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263, 2017.

[38] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang. Multi-modal fake news detection on social media via multi-grained information fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 343–352, 2023.

[39] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE, 2023.

[40] Y. Zhu, Q. Sheng, J. Cao, Q. Nan, K. Shu, M. Wu, J. Wang, and F. Zhuang. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 2022.