Unified Evidence Enhancement Inference Framework for Fake News Detection

Lianwei Wu\textsuperscript{1,2}, Linyong Wang\textsuperscript{1,2}, Yongqiang Zhao\textsuperscript{3}

\textsuperscript{1}ASGO, School of Computer Science, Northwestern Polytechnical University, Xi’an, China
\textsuperscript{2}Research & Development Institute of Northwestern Polytechnical University in Shenzhen
\textsuperscript{3}School of Computer Science, Peking University, Beijing, China

wlw@nwpu.edu.cn, linyongwang@mail.nwpu.edu.cn, yongqiangzhao@stu.pku.edu.cn

Abstract

The current approaches for fake news detection are mainly devoted to extracting candidate evidence from comments (or external articles) and establishing interactive reasoning with the news itself to verify the falsehood of the news. However, they still have several drawbacks: 1) The interaction object is coarse-grained, which mainly drives the entire news to participate in interaction, but ignores the learning of potential suspicious segments in news; 2) The reasoning ways are relatively single, making it difficult to explore the various possible correlations between news and candidate evidence. To this end, we propose Unified Evidence Enhancement Inference framework (UEEI) to discover and infer high-quality evidence for detection. Specifically, UEEI first promotes the interaction fusion between comments and news from the perspectives of semantics and emotion, thereby learning potential suspicious fragments in news. Then, the model constructs entity-level and relationship-level retrievals to screen sufficient candidate evidence from external sources. Finally, we measure coherence between suspicious fragments and candidate evidence by multi-view reasoning, and further infer explainable evidence. Experiments on three public datasets confirm the effectiveness and interpretability of our UEEI.

1 Introduction

In recent years, the growth and dissemination of fake news on social media have caused serious negative impacts. Especially in major events, such as the US elections [Grinberg \textit{et al.}, 2019], the Russo-Ukrainian War [Kreft \textit{et al.}, 2023], and the Israeli-Palestinian conflict [Kyriakidou \textit{et al.}, 2023], the deliberate and deceptive dissemination of fake news poses a significant threat to social stability and national security. In addition, the endless emergence of fake news makes it difficult for individuals to distinguish between truth and falsehood at short notice, posing a huge challenge to curb fake news [Van der Linden, 2023]. In light of these impacts and challenges, how to automatically detect fake news has become an imperative problem.

The current studies for fake news detection have made significant progress, which has gone through three stages. In addition to feature engineering stage of manually constructing features, with the popularization of deep learning, automatic fake news detection has greatly developed. Most methods design reasonable neural networks to extract features from the perspective of news content, which tends to learn semantic [Qian \textit{et al.}, 2021], emotional [Zhang \textit{et al.}, 2021], stance-based [Xie \textit{et al.}, 2021], and intentional [Zhou \textit{et al.}, 2022] features for detection. Besides, several methods pursue the patterns of news propagation, which establish graph neural networks [Xu \textit{et al.}, 2022] to explore the differences in propagation structures. Considering that the first two stages only give detection results but cannot supply interpretability, there is no evidence to prove news authenticity [Chien \textit{et al.}, 2022]. Interpretable detection receives extensive concern, which relies on acquiring knowledge from external sources and interacting with the news to capture corresponding evidence to reveal the errors of fake news [Guo \textit{et al.}, 2023].

Nevertheless, although these methods have improved the model performance, they still have several disadvantages: 1) They overlook the targeted exploration of potential suspicious fragments in news. The existing methods directly utilize the entire news content and interact with external evidence to uncover conflicting features between both. We know that the false parts in a news may only be hidden in a certain paragraph or sentence. The current methods are coarse-grained in terms of the entire news without specific screening, which makes it difficult to focus on suspicious fragments of the news; 2) The reasoning ways are relatively single. Existing methods generally devise simple attention interaction or semantic (or entity) alignment as inference strategies to discover consistency features between news content and potential evidence. These strategies mainly focus on global-level semantic similarity, without considering more possible relationships, which are easy to affect the accuracy of model inference.

To this end, we propose a Unified Evidence Enhancement Inference framework (henceforth, UEEI) to explore high-quality evidence from external sources and accurately locate the errors of the news for explainable detection. Specifically, in UEEI, 1) To explore potential suspicious fragments in news content, we develop hierarchical conflict discovery layer (HCD), which first learns key semantics from news and main viewpoints from comments, and then promotes interaction fusion from perspectives of semantic and emotions between key semantics and viewpoints, thereby fine-grained learning suspicious fragments in news; 2) To retrieve candidate evidence sentences...
from Wikipedia, external evidence enhancement layer (Ex3) is constructed, which designs different retrieval strategies from both entity-level and relationship-level, ensuring retrieval recall and accuracy, respectively; 3) Finally, we design multi-view coherence inference layer (MCI) that measures the consistency between suspicious fragments and candidate evidence from three aspects of causality, global, and local, which fully considers the possible relationships between the two, and further infers explainable evidence to confirm the false parts of the news. The experiments on three datasets demonstrate the superiority of UEEI. Our main contributions could be summarized as follows:

- A novel and unified evidence inference framework for explainable fake news detection is explored, which uncovers suspicious fragments within the news, obtains candidate evidence from external sources, and further explores valuable evidence to reveal the falsehood of fake news.

- HCD layer explores the interactive fusion between news content and comments from both semantic and emotional views, enabling fine-grained learning of suspicious fragments. Unlike traditional simple reasoning methods, MCI layer comprehensively considers multiple consistency relations and infers useful explainable evidence.

- Experimental results on three competitive datasets confirm that our UEEI achieves state-of-the-art performance and provides valuable and user-understandable evidence.

2 Related Work

Automatic Fake News Detection The existing methods mainly focus on extracting credibility-indicative features around news content and social context by constructing different neural networks. Content-based methods are inclined to learn features around semantics [Qian et al., 2021], emotions [Wan et al., 2023], writing styles [Zhou et al., 2023], and stances [Yang et al., 2022] from news content, relying on BiLSTM [Chen et al., 2023], attention mechanisms [Qian et al., 2021], and large language models [Hu et al., 2023]. Social context-based methods rely on graph neural networks to learn structural features of news propagation, thereby learning differences in propagation patterns among different news [Wei et al., 2022]. Dou et al. [2021] explored various signals from rich social contexts (e.g., users’ behavior history and social engagements) by joint content and graph modeling to demonstrate the effectiveness of propagation structures. In summary, the methods for automatic detection can deeply learn high-level representations in news content and social contexts to improve detection.

Explainable Fake News Detection The current methods [Wu et al., 2023a; Wang and Shu, 2023] mainly involve retrieving potential evidence from external sources and establishing different inference interaction mechanisms with the unverified news, so as to reveal the erroneous parts of the news. Different inference mechanisms mainly exploit similarity comparison [Yao et al., 2023], semantic matching [Wu et al., 2023a], entity alignment [Krishna et al., 2022], and consistency modeling [Wu et al., 2023c] strategies to respectively learn similarity semantics, common semantic fragments, associated entity information, and global consistency semantics between news and potential evidence. These methods have achieved a certain degree of interpretability, but they are still confronted with problems of ignoring the exploration of suspicious segments within the news and relatively single reasoning ways. Thus, we propose a novel evidence inference model that explores suspicious segments in news at a fine-grained level, and considers multiple consistency relationships between suspicious segments and candidate evidence to comprehensively learn effective evidence for detection.

3 The Proposed Model

3.1 Input Encoding Layer

The inputs of UEEI include three types: the title and content of the news, and all comments on the news. For any sequence with \( n \) words, it is represented as \( X = \{x_1, x_2, …, x_n\} \), where \( x_i \) is a \( d \)-dimensional vector obtained by pre-training BERT model aiming at the \( t \)-th word. For the encoding of each sequence \( X_t \), we adopt self-attention networks to capture contextual dependencies between words within a sequence, thereby learning the output \( E_c \). Here, the encoding sequences of title, content, and comments are denoted as \( E_T, E_C, \) and \( E_R \), respectively.

3.2 Hierarchical Conflict Discovery Layer (HCD)

With the continuous exposure of news, its comments are easily rich clues questioning the news credibility. To capture them, we construct HCD layer, which first relies on key semantic learning to obtain key semantics of news and main opinions in comments, and then establishes interaction between both semantic and emotion levels to explore questionable fragments.

Key Semantic Learning

Considering that some news is too long to capture valuable semantics, we build cross-attention networks to facilitate interaction between news title and content for capturing key semantics.

We utilize self-attention mechanism as cross-attention networks to capture the dependencies between any two tokens...
as to learn their contextual information, which is formalized as:

$$H = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$  \hspace{1cm} (1)

where $Q$, $K$, and $V$ are query, key, and value matrices, respectively. In our settings, we set $Q = E_T$ and $K = V = E_C$. $d_k$ is the column in comment encoding matrix. To enhance parallelism of the networks, we employ multi-head attention to linearly project queries, keys, and values:

$$h_{d_i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$ \hspace{1cm} (2)

$$H_F = \text{MultiH}(Q, K, V) = [h_{d_1}; h_{d_2}; ...; h_{d_m}]W_o$$ \hspace{1cm} (3)

where $W_i^Q$, $W_i^K$, $W_i^V$, and $W_o$ are all trainable parameters and $\phi$ is concatenation. $H_F$ is key fragments of the news.

We know that comments generally break down into several opinions. We design semantic-based cluster (i.e., single-pass incremental cluster) to learn mainstream viewpoints in all comments, which does not have any initial setting of clustering values, but determines whether to classify new comments into a new type by a similarity threshold $\theta_{sim}$. In this way, we obtain $K$ clusters with different viewpoints $H_R = \{H_{R1}, H_{R2}, ..., H_{RK}\}$.

**Hierarchical Interaction Fusion**

When encountering differences of opinions, people are prone to emotional conflicts. To identify more questionable fragments from comments, we establish hierarchical interaction between news and comments from semantic- and emotion-level.

**Semantic-level Interaction** To discover semantic contradictions, we employ cross attention to enhance interaction between key fragments of news and comment mainstream viewpoints:

$$H'_F = \text{Attention}(H_F, H_F, H_R)$$ \hspace{1cm} (4)

$$H'_{FR} = \text{Attention}(H_R, H_R, H_F), \quad H_{FR} = [H'_{F}; H'_{R}]$$ \hspace{1cm} (5)

where $H_{FR}$ is semantic-level questionable fragments.

**Emotion-level Interaction** The study [Zhang et al., 2022b] finds that news comments contain rich emotional words to deliver opinions. Thus, we devise emotion-level interaction to capture emotional questionable features between comments and news.

**Affective Graph Construction.** We build affective graph around news and comments separately. Specifically, given a news or comment sequence $X = \{x_1, x_2, ..., x_n\}$, we leverage sentiment dictionary SenticNet [Cambria et al., 2020] to evaluate emotional scores between any two words and obtain an adjacency matrix $E' \in \mathbb{R}^{n \times n}$, where each element $e_{ij}$ in $E'$ is:

$$e_{ij} = |u(x_i) - u(x_j)|, \quad u(x_i) \in [-1, 1]$$ \hspace{1cm} (6)

where $u(x_i)$ denotes the emotional score of the $i$-th word. $\cdot | \cdot$ is an absolute value operation. Especially, the greater the opposition between two emotional words, the higher their corresponding edge weights. In this way, words with conflicting emotions in the sequence will receive sufficient attention.

Additionally, given that emotional words with different contexts may convey different emotions (the context may contain different degree adverbs, negations, etc.), we exploit syntactic dependency trees (graph forms) to resolve the entire sequence to learn structural features between emotional words.

Each element in dependent adjacency matrix $D$ is:

$$d_{i,j} = \begin{cases} 
1, & v(i,j) = 1 \\
0, & v(i,j) = 0 
\end{cases}$$ \hspace{1cm} (7)

where $v(i,j) = 1$ indicates that there is an edge between words $x_i$ and $x_j$ in syntactic dependency tree. $v(i,j) = 0$ means there are no dependency relationships between two words.

To fuse sequence-structure and emotional features, we adopt coordination mechanism to control overall proportion between two graphs, thereby forming emotional enhancement graph.

$$E^{(l)} = \alpha E^{(l)} + (1 - \alpha)D^{(l)}$$ \hspace{1cm} (8)

where $\alpha$ is the hyperparameter. $E^{(l)}$ is the emotional enhancement graph obtained after $l$ iterations.

**Aggregator Fusion.** Naturally, emotional information in a sequence is more discrete than semantic features, which may lead to sparse issue in the construction of affective graph. To alleviate it and enhance deep fusion between emotional and semantic-level questionable information, we design two aggregator fusion modes:

1) **Attention Aggregator Fusion.** We use semantic-level questionable information to attend to aggregation of affective graph for exploring common questionable features $O_a$.

$$\alpha_i = W_o(\text{ReLU}(W'_o(H_{FR}|E^{(l)}_o)))$$ \hspace{1cm} (9)

$$\beta_i = \text{softmax}(\alpha_i) = \frac{\exp(\alpha_i)}{\sum_{k=1}^{N}\exp(\alpha_k)}$$ \hspace{1cm} (10)

$$O_a = \sum_{k=1}^{N}\exp(\beta_i E^{(l)}_k)$$ \hspace{1cm} (11)

where $W_o$ and $W'_o$ are trainable parameters. $N$ is the number of the nodes of affective graph.

2) **Max Aggregator Fusion.** Questionable information is more likely to appear in extreme emotional expressions. Therefore, we also explore element-wise maximization to aggregate significant emotion information, which adopts dynamic weight to balance proportion between semantic- and emotion-level features, and then maximizes the mining of major emotionally questionable features.

$$\mu_m = \sigma(H_{FR}W_{\mu1} + E^{(l)}W_{\mu2} + W_{\mu})$$ \hspace{1cm} (12)

$$O_a = \text{Tanh}(\mu_m \text{Max}(E^{(l)}_1, E^{(l)}_2, ..., E^{(l)}_m)W_m + (1 - \mu_m)E^{(l)}W_n + b_m)$$ \hspace{1cm} (13)

where all $W$ and $b$ are trainable parameters. Finally, we integrate the two fusion modes to maximize questionable information between news and comments, i.e., $O = [O_a; O_m]$.

**3.3 External Evidence Enhancement Layer (Ex3)**

We first design dual-level keyword retrieval to respectively improve retrieval accuracy and recall, to comprehensively capture relevant articles. Then, evidence selection screens fine-grained sentences from articles as external evidence.

**Dual-level Keyword Retrieval**

**Entity-level Keyword Retrieval** Entity-level keyword retrieval endeavors to collect and optimize retrieval keyword seeds, which could retrieve wide-coverage results and ensure retrieval recall.

1) **Keyword Matching:** 1) Title-guided matching: We employ news title as keyword seeds for retrieval from Wikipedia.

2) **Co-occurrence words extracting:** Due to title and content as
Figure 2: Three coherence modeling in the MCI layer of UEEI.

a whole, we extract their co-occurrence words and sort them by quantity, and choose the top-K words as seeds for retrieval.

Step 2: Keyword Seed Optimization: If there are few feedback results in step 1, keyword seed optimization will be carried out by removing stop words and filtering keywords with low importance in sequence. Filtering way relies on TF-IDF to filter out the least important words for keyword optimization.

Relationship-level Keyword Retrieval It endeavors to boost retrieval accuracy by extracting correlation words in news (like subject-predicate), which are more effective in depicting news opinions and making it easier to retrieve highly relevant articles, so they are more inclined to improve retrieval accuracy.

Inspired by the work [Zhong et al., 2020], we use semantic role labeling (SRL) [Carreras and M`arquez, 2005] to parse news titles and content, establish connections between arguments to construct graphs, and then exploit graph-based reasoning methods to obtain keyword seeds in news, which contain rich association relationships between key semantics. These seeds can be used to search for more accurate related articles. Finally, we integrate the articles retrieved from the two levels to form relevant article set A.

Evidence Selection
This block aims to extract evidence sentences \( E_c \) from A.

\[
G(X, U, P_s) = E_c
\]

where \( E_c \) is the i-th sentence of \( E_c \). Evidence selection could be regarded as a semantic matching task, where each sentence in an article is compared with a given news to determine the probability of that sentence becoming evidence. Here, we utilize BART [Lewis et al., 2020] to improve sentence selection. It inputs news X and sentences S of a retrieved article into model in the following way: [CLS]X[SEP]S[SEP][EOS]. For outputs, following the work [Liao et al., 2023], we employ the BIO form to classify irrelevant tokens as O, where the first token in evidence sentence is set as B evidence, and the rest tokens of a sentence as I evidence. We perform the above operations on all articles, and ultimately obtain the most likely evidence set with p sentences as \( E = \{e_1, e_2, \ldots, e_p\} \).

3.4 Multi-view Coherence Inference Layer (MCI)
We propose MCI layer that includes multiple coherence modeling and synthetic inference. The former learns coherence relationships of internal conflicts between news and external evidence to discover different incoherence semantics, and the latter is devoted to fine-grained inferring explainable evidence.

Multiple Coherence Modeling
To reveal the inconsistency between conflicting semantics of news and external evidence, multiple coherence modeling promotes deep interaction between both. It mainly includes causal coherence, mutual coherence, and refined coherence, which respectively attempt to explore incoherence features of different perspectives: causal, global, and local hotspots.

Three coherence blocks are all extended self-attention mechanism (Eqs. 15 and 16), which focuses on adjusting nonlinear transformation \( f(\cdot) \) to adapt to different coherence strategies.

\[
att_i = \text{softmax}(v_i^T f(h_i))
\]

\[
r = \sum_{i=1}^{n} att_i h_i
\]

Causal Coherence. It primarily maps the conflicting information of news and external potential evidence into the same dimensional space, and then conducts interactive comparisons to explore the causal relationship between them.

\[
\alpha'(O) = W_o O + b_o, \quad \beta(E) = W_e E + b_e, \quad \gamma(E) = W_e E + b_e
\]

\[
G_c = \ell_c(O, E) = \sigma(\beta(E)) \odot \alpha'(O) + \gamma(E)
\]

where all W and b are parameters and \( \odot \) is element-wise multiplication.

Mutual Coherence. To discover the incoherence features between conflicting information and potential evidence from a global perspective, we design mutual coherence that maps and transforms them together to learn their inconsistent features:

\[
G_m = f_m(O, E) = \text{tanh}(W_m O|E| + b_m)
\]

Refined Coherence. For highly focused conflicting information, we explore refined coherence, which first highly condenses the conflicting information in the news, and then further interacts with external evidence to refine their inconsistent features:

\[
G_r = f_r(O, E) = \sigma(W_r E + b_r) \odot O
\]

where \( W_r \) and \( b_r \) are trainable parameters. Thus, the above is integrated as multi-view consistent semantics \( G \) between conflicting information and external evidence: \( G = G_c + G_m + G_r \).

Synthetic Inference
To further synthesize explainable evidence, we propose synthetic inference to drive contextual association between multi-view consistent semantics and external evidence.

Cross-Attention Networks. It explores common false parts of news \( H_g \) and highly-correlated evidence \( H_e \) respectively by deep interaction association of different semantics.

\[
H_g = \text{Attention}(G, G, O), \quad H_e = \text{Attention}(G, G, E)
\]

Heuristic Fusion. It establishes full integration between common false parts and highly-correlation evidence, thereby compounding accurately interpretable evidence.

\[
H_{gge} = \text{max}(H_g, H_e) \odot \text{max}(H_g, H_e)
\]

Next, we apply feed forward network to the fusion results for adding non-linear features while scale-invariant features, which involves a single hidden layer with an ReLU.

\[
l_g = \text{FFN}(H_{gge}) = \text{max}(0, H_{gge})W_1 + b_1W_2 + b_2
\]

\[
l_e = \text{FFN}(H_{gge}) = \text{max}(0, H_{gge})W_3 + b_3W_4 + b_4
\]

Finally, we adopt softmax function to emit probability distribution for training, where a loss drives the model to minimize
We compare UEEI with the following baselines: TextRNN, Weibo platform, which mainly involves some hot news and website for factual checks of celebrity reports. The labels are annotations for detection. [Yang et al., 2016] employs LSTM to learn contextual representations for detection. TCNN-URG [Qian et al., 2018] integrates user responses into two-level CNN to learn contextual semantics to detect fake news. BERT [Kenton and Toutanova, 2019] focuses on gaining long-term dependencies within news for detection. DeClarE [Popat et al., 2018] explores potential evidence between claims and articles for debunking claims. HAN [Ma et al., 2019] is hierarchical attention networks, which learns coherence between news and related articles to obtain interpretability. EHIAN [Wu et al., 2021] builds inference networks to identify semantic conflicts from related articles as evidence. MAC [Vo and Lee, 2021] leverages dual-view attention to explore word-level and document-level evidence for detection. GET [Xu et al., 2022] develops graph networks to obtain semantic structures and model evidence to detect fake news. MUSER [Liao et al., 2023] builds multi-step retrieval and learns dependencies between pieces of evidence for detection.

### 4 Experiments

In this section, we endeavor to answer the following questions:

- **Q1:** Could UEEI achieve more excellent performance?
- **Q2:** Does each layer contribute to improving detection?
- **Q3:** How much does the exploration of potentially suspicious fragments of news boost model performance?
- **Q4:** What are the advantages of our multi-view coherence inference compared to existing reasoning ways?
- **Q5:** Is the obtained evidence reasonable and interpretable?

#### 4.1 Datasets and Evaluation Metrics

As shown in Table 1, PolitiFact and GossipCop are two English datasets [Shu et al., 2020] and Weibo is a Chinese dataset [Liu et al., 2018]. PolitiFact is a factual verification website that deals with political topics and rating credibility of voter officials’ claims. This dataset mainly includes two types of labels: real and fake. The labels are evaluated by professional journalists. GossipCop is collected from the Gossip Cop website for factual checks of celebrity reports. The labels are the same as PolitiFact dataset. Weibo is collected from Sina Weibo platform, which mainly involves some hot news and is marked as true news and rumors. **Evaluation Metrics.** We leverage F1-macro (F1-Ma), and F1-micro (F1-Mi), and F1 to measure our model. In dataset partitioning, we hold out 75% of the news as training set and the remaining 25% as test set.

#### 4.2 Settings

For preprocessing, we tokenize sequences and convert tokens to lowercase, filtering out nonalphabetical characters. For model configuration, we adopt BERT-base as embeddings. The embedding size $d$ is 768. $K$ clusters of viewpoints are from [3, 4, 5], and $\alpha$ is 0.65. $K$ in the top-$K$ words is 10; In self-attention networks, attention heads and blocks are set to 6 and 4, respectively, and the dropout of multi-head attention is 0.6. We adopt Adam optimizer as the model optimizer. The learning rate is uniformly set to $10^{-4}$. We utilize L2-regularizers with the fully-connected layers and the mini-batch size is 32.

#### 4.3 Performance Comparison (Q1)

**Comparative Baselines**


#### 4.4 Discussion

#### 4.5 Ablation Study (Q2)

To assess the necessity of each module in our model, we ablate UEEI into: -HCD, -Ex3, and -MCI are the removal of HCD, Ex3, and MCI layers. -Ex3(entity), -Ex3(relation), and -Ex3(select) are denoted as entity-level retrieval, relationship-level retrieval, and evidence selection blocks ablated from UEEI, respectively. As shown in Table 3, we could observe that:

- The separation of MCI results in a significant change, with up to 5.2% degradation on three datasets, which sufficiently demonstrates the value of establishing consistency reasoning from the perspectives of causality, global, and local.
- In Ex3 layer, the removal of entity- and relationship-level retrieval blocks results in varying degrees of model degradation, which confirms respectively the effectiveness of both boosting retrieval recall and accuracy through entity and relationship aspects. -Ex3(select) is weaker than our model, which reveals the rationality of extracting key evidence sentences from documents.
- Overall, the ablation of each module weakens the overall performance of the model, showing a decrease from 1.4% to 5.2% in F1-Ma on three datasets, which reveals the effectiveness of various modules that constitute our model.

<table>
<thead>
<tr>
<th></th>
<th>PolitiFact</th>
<th>GossipCop</th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td>#True News</td>
<td>399</td>
<td>4,219</td>
<td>436</td>
</tr>
<tr>
<td>#Fake News</td>
<td>345</td>
<td>3,393</td>
<td>311</td>
</tr>
<tr>
<td>Total</td>
<td>744</td>
<td>7,612</td>
<td>747</td>
</tr>
</tbody>
</table>

Table 1: Statistics of three publicly available datasets.

$$\text{Loss} = -\sum y \log p$$

$$p = \text{softmax}(W_p I_{geo} + b_p)$$

**Overall Performance**

As shown in Table 2, we observe that:

- In the methods of automatic feature learning (i.e., the first three baselines), TCNN-URG achieves the optimal performance, indicating that relying on comments as external features can markedly improve model performance.
- Compared to automatic methods, the methods of exploring evidence (like DeClarE) achieve better performance, showing the improvements from 0.8% to 4.9% in F1-Mi on the three datasets. This reflects the fact that capturing conflicting semantics between news and external articles as credibility indicators is conducive to facilitating detection.
- Our UEEI consistently outperforms all baselines, showing up to 3.4% improvements in all metrics on three datasets than the latest baseline (MUSER). This not only highlights the superiority of our model, but also demonstrates that the route of exploring fine-grained error parts of the unverified news and capturing high-quality evidence is effective.
Table 2: Performance comparison of UEEI against the baselines on the three datasets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Politifact</th>
<th>GossipCop</th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1-Ma</td>
<td>F1-Mi</td>
<td>F1-T</td>
</tr>
<tr>
<td>TextRNN</td>
<td>61.0</td>
<td>60.9</td>
<td>61.6</td>
</tr>
<tr>
<td>TCNN-URG</td>
<td>62.1</td>
<td>61.9</td>
<td>63.7</td>
</tr>
<tr>
<td>BERT</td>
<td>59.7</td>
<td>59.8</td>
<td>60.8</td>
</tr>
<tr>
<td>DeClarE</td>
<td>65.4</td>
<td>65.1</td>
<td>65.6</td>
</tr>
<tr>
<td>HAN</td>
<td>66.1</td>
<td>66.0</td>
<td>67.9</td>
</tr>
<tr>
<td>EHIAN</td>
<td>66.4</td>
<td>66.3</td>
<td>67.4</td>
</tr>
<tr>
<td>MAC</td>
<td>67.8</td>
<td>67.5</td>
<td>70.0</td>
</tr>
<tr>
<td>GET</td>
<td>69.4</td>
<td>69.2</td>
<td>72.5</td>
</tr>
<tr>
<td>MUSER</td>
<td>73.2</td>
<td>72.9</td>
<td>75.7</td>
</tr>
</tbody>
</table>

UEEI(Ours) 76.4 75.6 78.9 73.6 80.4 80.1 81.0 79.2 82.7 82.7 85.0 81.7

Table 3: Results of ablation analysis of our UEEI on the three datasets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Politifact</th>
<th>GossipCop</th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1-Ma</td>
<td>F1-Mi</td>
<td>F1-T</td>
</tr>
<tr>
<td>-HCD</td>
<td>0.729</td>
<td>0.715</td>
<td>0.755</td>
</tr>
<tr>
<td>-Ex3</td>
<td>0.713</td>
<td>0.708</td>
<td>0.741</td>
</tr>
<tr>
<td>-MCI</td>
<td>0.718</td>
<td>0.704</td>
<td>0.746</td>
</tr>
<tr>
<td>-Ex3(entity)</td>
<td>0.745</td>
<td>0.740</td>
<td>0.775</td>
</tr>
<tr>
<td>-Ex3(relation)</td>
<td>0.736</td>
<td>0.734</td>
<td>0.757</td>
</tr>
<tr>
<td>-Ex3(select)</td>
<td>0.741</td>
<td>0.737</td>
<td>0.771</td>
</tr>
<tr>
<td>UEEI</td>
<td>0.764</td>
<td>0.756</td>
<td>0.789</td>
</tr>
</tbody>
</table>

Evaluation of Hierarchical Conflict Discovery (Q3)

To analyze the effectiveness of the model's fine-grained learning of suspicious fragments in news, we explore the contribution of each block in HCD layer. We divide contributions into two categories: semantic-based and emotion-based. Semantic-based contribution: -keyFrag, -mainOpinion, -HIF(NewsCA), -HIF(CommCA) represent respectively that the model replaces news key fragments, main opinions in comments, as well as cross-attention aiming to news content and comments in hierarchical interaction fusion as concatenation operations. Emotion-based contribution: -NewsAG and -CommAG are the replacement of affective graphs for news and comments by BiLSTM, respectively. As shown in Figure 3, we observe that:

- The detachment of modules related to suspicious fragments induces varying degrees of degradation. Especially, the removal of clustering algorithm aiming at comments has been greatly affected. These reflect the superiority of capturing suspicious fragments and utilizing comments.
- The alteration of affective graph contributes to a decrease of at least 0.6% in F1-Mi, indicating that using graph neural networks (GNN) is more effective in learning emotions than sequence models (BiLSTM). The reason is that emotional words in news or comments are discrete, and they are not just simple context relationships, and we can extract complex long-range correlations between emotion words using GNN.
- The changes in both types of modules cause a reduction in performance, which illustrates the effectiveness of each module and mutual promotion of semantic and emotional layers in extracting suspicious fragments in the news.

Superiority of Multi-view Coherence Inference Layer (Q4)

To further validate the superiority of MCI layer, we replace it with advanced inference modules: ECA_infer [Zhang et al., 2022a] is that two extended attentions with feed-forward networks perform feature- and relation-level inference. DocInfer
Model Validation with User Participation
To validate the comprehensibility of the evidence, we study from two perspectives: 1) Users rely on candidate evidence captured by UEEI and the latest retrieval way (MR) [Liao et al., 2023] to rate the unverified news for verifying the two types of evidence; 2) Adopting user participation to manually select more favorable evidence from candidate evidence to downstream for measuring the effectiveness of man-in-the-loop. In detail, we randomly choose 120 news from three datasets and retrieve Top-5 evidence articles, with 5 real-world users as participants. As shown in Figures 6 and 7, we observe that: From Figure 6, compared to MR, users gain better performance in classifying news based on our captured evidence, showing at least a 2.7% boost in F1-Mi, and our user satisfaction degree is significantly higher than that of MR, which demonstrates that our retrieved evidence is easier to understand and interpretable. From Figure 7, evidence manually screened by users outperforms evidence automatically retrieved by UEEI, with improvements of up to 0.7% on the three datasets, which confirms the effectiveness of user participation in detecting fake news. Furthermore, human-in-the-loop strategy is time-consuming and labor-intensive, and we need to carefully consider the combination of automatic detection and user participation.

5 Conclusion
In this paper, we propose unified evidence enhancement inference model (UEEI) for fake news detection based on the progressive process of suspicious fragment learning, evidence enhancement, and coherence inference. We first promote hierarchical interaction between news and comments to explore internal suspicious fragments, then enhance external evidence retrieval around two levels, and finally drive multi-view coherence learning between suspicious semantics and external evidence, thereby inferring false parts of news. Experiments on three datasets demonstrate the effectiveness and interpretability of our UEEI. In the future, considering current prevalence of multimodal news, we seek to design inference mechanisms to explore inconsistency between multimodal news for detection.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grants U22B2036 and 62202381, in part by Shenzhen Science and Technology Program and Guangdong Basic and Applied Basic Research Foundation (2021A1515110717, 2024A1515010087), General Program of Chongqing Natural Science Foundation (No. CSTB2022NSCQ-MSX1284). Sponsored by CAAI-Huawei MindSpore Open Fund, the National Postdoctoral Innovative Talents Support Program for L. Wu. We would like to thank the anonymous reviewers for their constructive comments.

References


[Qian et al., 2021] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection.


