A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection

Zhiwei Yang[1,2,3,5], **Jing Ma**[2,*], **Hechang Chen**[3,5,*], **Hongzhan Lin**[2], **Ziyang Luo**[2], **Yi Chang**[3,4,5,*]

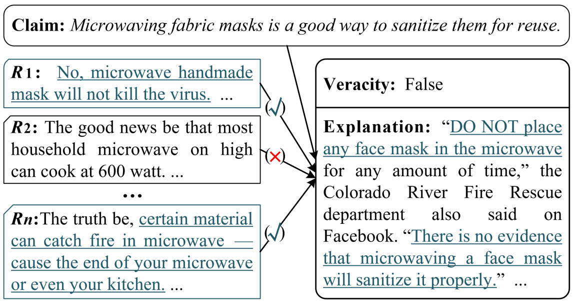
[1] College of Computer Science and Technology, Jilin University, Changchun, China
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[3] School of Artificial Intelligence, [4] International Center of Future Science, Jilin University, China
[5] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education
yangzwl8@mails.jlu.edu.cn, chenhc@jlu.edu.cn,
{majing, cszyluo, cshzlin}@comp.hkbu.edu.hk, yichang@jlu.edu.cn

Abstract

Existing fake news detection methods aim to classify a piece of news as true or false and provide veracity explanations, achieving remarkable performances. However, they often tailor automated solutions on manual fact-checked reports, suffering from limited news coverage and debunking delays. When a piece of news has not yet been fact-checked or debunked, certain amounts of relevant raw reports are usually disseminated on various media outlets, containing the wisdom of crowds to verify the news claim and explain its verdict. In this paper, we propose a novel Coarse-to-fine Cascaded Evidence-Distillation (CofCED) neural network for explainable fake news detection based on such raw reports, alleviating the dependency on fact-checked ones. Specifically, we first utilize a hierarchical encoder for web text representation, and then develop two cascaded selectors to select the most explainable sentences for verdicts on top of the selected top-K reports in a coarse-to-fine manner. Besides, we construct two explainable fake news datasets, which is publicly available. Experimental results demonstrate that our model significantly outperforms state-of-the-art detection baselines and generates high-quality explanations from diverse evaluation perspectives.

1 Introduction

During the COVID-19 pandemic, almost 80% of consumers in the United States received fake news, which has caused confusion and undermined public health efforts[1]. The proliferation of fake news has increased the demand for automatic fake news detection (Guo et al., 2022). To further clarify and explain detection results, explainable fake news detection has gained more importance recently, aiming to classify the truthfulness of a piece of news and generate veracity explanations[2] (Kotonya and Toni, 2020a). However, existing methods have a limitation in detecting fake news timely as they heavily relied on debunked reports of investigated journalism. Thus, it is urgent to develop explainable yet general methods to mitigate this issue.



Claim: *Microwaving fabric masks is a good way to sanitize them for reuse.*

***R*1:** No, microwave handmade mask will not kill the virus. ...

***R*2:** The good news be that most household microwave on high can cook at 600 watt. ...

...

***Rn*:** The truth be, certain material can catch fire in microwave — cause the end of your microwave or even your kitchen. ...

Veracity: False

Explanation: "DO NOT place any face mask in the microwave for any amount of time," the Colorado River Fire Rescue department also said on Facebook. "There is no evidence that microwaving a face mask will sanitize it properly." ...

Figure 1: An example for veracity explanation generation. The underlined explanations can be semantically inferred from some relevant sentences in the reports R_1 and R_n. "R" denotes the raw report.

Many previous approaches detected fake news without any justifications (Wang, 2017; Ma et al., 2018). Recently, some explainable methods highlighted salient words or phrases in relevant reports as explanations (Popat et al., 2018; Wu et al., 2021), which lack readable complete sentences. To alleviate these issues, some methods aimed to extract salient sentences from relevant reports via attention mechanisms (Nie et al., 2019; Ma et al., 2019), or pre-trained extractive-abstractive summarization (Kotonya and Toni, 2020b), etc. As the human justification about veracity labels can significantly improve the performance of veracity prediction (Alhindi et al., 2018), Atanasova et al. (2020) proposed the first study on producing veracity explanations jointly with veracity prediction utilizing the debunked report released by fact-checking websites. However, such a debunked report is based on manual endeavors, thus prone to be coverage-limited and relatively inefficient.

A new study by MIT researchers suggests that crowds of laypeople reliably rate claims as effectively as fact-checkers do (Allen et al., 2021). To

[*] Corresponding authors.

[1] https://www.statista.com/topics/3251/fake-news

[2] Explanations and evidence are used interchangeably

Proceedings of the 29th International Conference on Computational Linguistics, pages 2608–2621
October 12–17, 2022.

use the wisdom of crowds, we assume that crowds of relevant raw reports (e.g., media reports, user comments, blogs, etc.) published by different media outlets contain evidence for effectively detecting fake news and explaining verdicts (Ma et al., 2019; Popat et al., 2018). As shown in Figure 1, given a false claim "*Microwaving fabric masks is a good way to sanitize them for reuse*", the check-worthy reports R_1 and R_n are selected from all reports $[R_1, R_2, \cdots, R_n]$ and then some evidential sentences (underlined) can be used to generate veracity explanations. In contrast, existing methods usually tailor models on one manual fact-checked article, rarely attempting to detect fake news based on raw reports.

To this end, we propose a general coarse-to-fine cascaded evidence-distillation (CofCED) network to detect fake news and explain verdicts directly using raw reports, mitigating the dependency on fact-checked reports. Specifically, we design a hierarchical encoder for text representation, and then we develop two coarse-to-fine cascaded selectors to distill explainable sentences on top of the selected top-K check-worthy reports. Our predictions of explainable sentences can be obtained by explicitly considering four features, i.e., claim relevance, richness, salience, and non-redundancy. Different from FEVER (Thorne et al., 2018) using human-crafted claims with credible Wikipedia articles, the claims in our task are real-world news containing some unreliable reports. Thus, detecting fake news on raw reports is much more challenging and significant than that in FEVER task.

Our contributions are as follows: **1)** To the best of our knowledge, we present the first study on explainable fake news detection directly utilizing the wisdom of crowds, alleviating the dependency on fact-checked reports; **2)** Our model has the advantage of revealing insight into the generation of veracity explanations from various perspectives; **3)** We construct two realistic datasets, i.e., RAWFC and LIAR-RAW, consisting of raw reports for each claim. Experimental results on benchmarks demonstrate the effectiveness of CofCED for detecting fake news and and explaining verdicts based on raw reports. Our resources are publicly available at https://github.com/Nicozwy/CofCED.

2 Related Work

We review prior works closely related to ours based on several surveys (Shu et al., 2017; Kotonya and Toni, 2020a).

Black-boxed fake news detection. Many existing studies on fake news detection achieved promising performances by incorporating claim metadata to facilitate the detection, such as user profiles (Wang, 2017; Long, 2017; Karimi et al., 2018). Besides, various deep learning methods have been proposed to capture report features, e.g., credibility (Popat et al., 2017), stances (Ma et al., 2018), writing styles (Potthast et al., 2018), extra knowledge (Dun et al., 2021), etc. Although these methods could improve the detection performance, they are lack of explainability on verdicts.

Explainable fake news detection. To address the above issue, many explainable methods on this task explored attention mechanisms to highlight salient words (Popat et al., 2018; Wu et al., 2021), news attributes (Yang et al., 2019), and suspicious users (Lu and Li, 2020), to obtain relevant evidence, providing a certain explainability. To improve the readability in word-level methods, there are some methods obtained evidential sentences using attention weights (Shu et al., 2019), semantic matching (Nie et al., 2019), and entailment (Ma et al., 2019). More recently, Atanasova et al. (2020) proposed the first study on directly producing veracity explanations using extractive summarization, and Kotonya and Toni (2020b) made use of extractive-abstractive summarization for explanation generation, independent of the veracity prediction. However, they significantly relied on the manual fact-checked report and rarely attempted to consider fine-grained features for this task. Thus, we utilize the wisdom of crowds for fake news detection based on raw reports, providing a highly explainable structure for explanation generation.

Datasets. For explainable fake news detection, FEVER (Thorne et al., 2018) was crafted merely from credible Wikipedia articles, and MultiFC (Augenstein et al., 2019) provided a real-world benchmark for multi-domain claims. While offering evidence labels, they do not contain veracity explanations. By contrast, LIAR-PLUS (Alhindi et al., 2018) extended on LIAR (Wang, 2017) and PUBHEALTH (Kotonya and Toni, 2020b) on the public health, providing manual explanations for explainable fake news detection. However, they only contain the manual fact-checked report that is relatively inefficient and coverage-limited. Thus, we constructed two datasets by collecting raw reports, which is more suitable and challenging for this task.

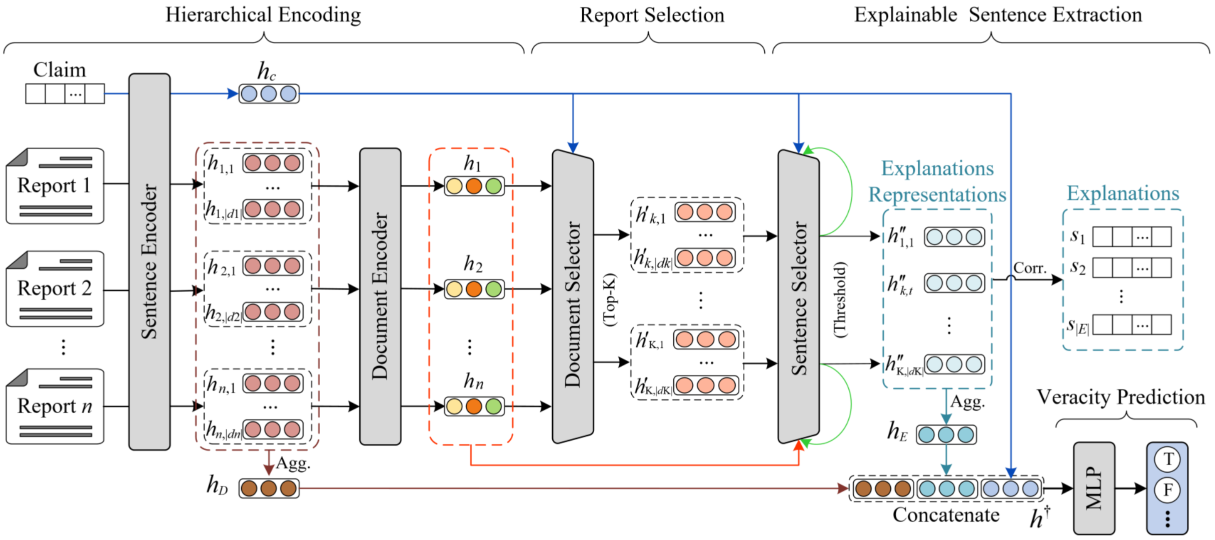


Figure 2: An overview of our proposed CofCED framework. The document selector and the sentence selector are used for selecting check-worthy reports (containing oracles) and oracles, respectively. “Agg.” denotes aggregation and “ Corr.” denotes corresponding. We use different color to highlight different objects. Note that the green line denotes the last output of sentence selection for checking redundancy.

3 Problem Statement

Given a fake news dataset $\{C\}$, $C = (c, \mathcal{D})$ is a tuple representing a given claim c and its relevant raw reports $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$, where each $d_i = (s_{i,1}, s_{i,2}, \cdots, s_{i,|d_i|})$ denotes a relevant report consisted of a sequence of sentences and $|.|$ denotes the number of items. In the task of explainable fake news detection, each claim c is associated with a veracity y taking one of the class labels from $\{\text{True}, \text{False}, \cdots\}$, and each raw report d_i is associated with a binary label $y_i^d \in Y^d$ indicating that whether d_i contains explainable sentences (i.e., oracles). For each sentence $s_{i,j}$, $y_{i,j}^s \in Y^s$ is a binary label indicating that whether $s_{i,j}$ is one of the explainable sentences w.r.t. the gold justification.

We formulate this task as a multi-task learning problem by considering check-worthy report selection, explainable sentence extraction, and veracity prediction. Formally, $f : f(c, \mathcal{D}) \rightarrow (\hat{y}, \hat{Y}^d, \hat{Y}^s, \hat{E})$, where $\hat{E}$ denotes the veracity explanation (i.e., *evidence*) consisting of a set of predicted sentences (i.e., $\hat{y}_i^d = 1$ and $\hat{y}_{i,j}^s = 1$).

4 CofCED: The Proposed Method

Fig. 2 gives an overview of our proposed CofCED, which consists of four parts: hierarchical encoding, report selection, explainable sentence extraction, and veracity prediction.

4.1 Hierarchical Encoding

Given a word sequence of a claim or report sentence $T = (w_1 \cdots w_t \cdots w_{|T|})$, where $w_t \in \mathbb{R}^d$ is a d-dimensional vector initialized with a text encoder. Because words form a sentence and sentences form a report, we utilize a hierarchical encoding method for sentence and report representation in our model. Specifically, for sentence encoding, we use the special token “[CLS]” embedding from the final contextual layer of the pre-trained language model (Sanh et al., 2019) as the sentence representation. Thus, we obtain the sentence representation for a claim c and each sentence $s_{i,j}$ in a raw report d_i as $\mathbf{h}_c \in \mathbb{R}^d$ and $\mathbf{h}_{i,j} \in \mathbb{R}^d$, respectively.

For document encoding, we further adopt a document encoder consisting of a bidirectional LSTM (BiLSTM) (Rashkin et al., 2017) and a max-pooling layer to aggregate all salient sentence features as the representation of a report:

$$\tilde{\mathbf{h}}_{i,j} = \text{BiLSTM}(\mathbf{h}_{i,j}, \overrightarrow{\mathbf{h}}_{i,j-1}, \overleftarrow{\mathbf{h}}_{i,j-1}, \theta) \qquad (1)$$

$$\mathbf{h}_i = \text{Max}([\tilde{\mathbf{h}}_{i,1}; \tilde{\mathbf{h}}_{i,2}; \cdots; \tilde{\mathbf{h}}_{i,|d_i|}]) \qquad (2)$$

where $\tilde{\mathbf{h}}_{i,j} \in \mathbb{R}^d$ denotes the cross-sentence hidden state, and $\mathbf{h}_i \in \mathbb{R}^d$ denotes the representation of the report d_i. Max denotes the max pooling, [;] denotes concatenation, and θ denotes encoder parameters.

4.2 Report Selection

Since this task is formulated on massive raw reports, our model aims to automatically narrow down the evidence extraction by ranking them and capturing the top ones for further analysis. Taking the claim in Fig. 1 as an example, there are n retrieved reports about "microwaving fabric masks" and the significant reports R_1 and R_n containing oracles (i.e., underlined sentences) are selected for veracity prediction and explanation generation.

To distill the check-worthy reports from massive reports $\mathcal{D}$ that are helpful for veracity prediction, we firstly develop a coarse-grained document selector by treating the claim as a query to find K most significant results. Then, global attention is utilized to obtain the significance score for each report d_i:

$$\alpha_{c\rightarrow\mathcal{D}} = \mathrm{softmax}(\mathbf{H}_{\mathcal{D}} W_\alpha \mathbf{h}_c) \tag{3}$$

where $\mathbf{H}_{\mathcal{D}} = [\mathbf{h}_1; \mathbf{h}_2; \cdots; \mathbf{h}_{|\mathcal{D}|}]$ compacts all hidden vectors of reports and $W_\alpha \in \mathbb{R}^{d\times d}$ is a trainable parameter. We use $\alpha_{c\rightarrow\mathcal{D}}$ to rank all reports and select the top-K results as the check-worthy reports (i.e., $\hat{y}_i^d = \alpha_i(\alpha_i \geq \alpha_K)$ and otherwise $\hat{y}_i^d = 0(\alpha_i < \alpha_K)$). Note that the t-th sentence representation in the k-th selected report $d_k^{'}$ are denoted as $\mathbf{h}_{k,t}^{'} \in \{\mathbf{h}_{k,1}^{'}, \mathbf{h}_{i,2}^{'}, ..., \mathbf{h}_{k,|d_k^{'}|}^{'}\}$, and its document representation is denoted as $\mathbf{h}_k^{'}$, which are used for explainable sentence extraction.

4.3 Explainable Sentence Extraction

On top of selected reports, we treat explanation generation as a multi-document extractive summarization, where each report is visited sequentially for explainable sentences. Such reports are regarded as the wisdom of crowds when detecting a dubious claim. We assume that explainable sentences for verdicts should be claim-relevant, informative, salient, and non-redundant. Specifically, there may exist redundancy between reports because a report is generally self-contained and multiple raw reports are more likely to contain semantically irrelevant and redundant sentences (Ma et al., 2019).

In this paper, we develop a fine-grained sentence selector to extract explainable sentences from these check-worthy reports considering the following four features: 1) *claim relevance* measures the topic coverage of each sentence regarding the claim; 2) *richness* measures the content informativeness of each sentence containing evidence; 3) *salience* measures the significance of each sentence regarding the entire report; 4) *non-redundancy* measures the novelty of each sentence regarding previous selected explainable sentences. Therefore, we define a layer to predict the probability of each sentence that should be selected via integrating the four features as follows:

$$\begin{aligned}
\mathrm{P}(y_{k,t}^s = 1 | \mathbf{h}_c, \mathbf{h}_{k,t}^{'}, \mathbf{h}_k^{'}, \mathbf{h}_d) \\
= \sigma(\underbrace{\mathbf{h}_{k,t}^{'} W_c \mathbf{h}_c}_{(\textit{claim relevance})} + \underbrace{\mathbf{h}_{k,t}^{'} W_s}_{(\textit{richness})} \\
+ \underbrace{\mathbf{h}_{k,t}^{'} W_r \mathbf{h}_k^{'}}_{(\textit{salience})} - \underbrace{\mathbf{h}_{k,t}^{'} W_d \mathbf{h}_d}_{(\textit{non-redundancy})})
\end{aligned} \tag{4}$$

where $y_{k,t}^s$ is a binary variable indicating whether the t-th sentence in the selected report $d_k^{'}$ should be selected as part of explanations $\hat{E}$, and W_* are trainable parameters. $\mathbf{h}_d$ is the redundancy vectors initialized with all zeros and updated by selected sentences in previously visited reports as follows:

$$\mathbf{h}_d = \tanh(\sum_t \mathbf{h}_{k-1,t}^{'} \cdot \mathrm{P}(y_{k,t}^s = 1)) \tag{5}$$

Considering the number of report sentences, our model learns to select the explainable sentences with probabilities above a soft threshold $\varepsilon_k = 1/|d_k^{'}|$, i.e., $\mathrm{P}(y_{k,t}^s = 1) > \varepsilon_k$, where $\mathrm{P}(y_{k,t}^s = 1)$ is obtained by Eq. (4). Note that $\mathbf{h}_{k,t}^{''}$ is used to denote the sentence representation output from the explainable sentence selector.

4.4 Veracity Prediction

To enhance final veracity prediction, we further employ the extracted explanation as additional *evidence* besides the claims and all reports. Specifically, we aggregate the recognitions from such evidence and reports for a target claim, respectively, and then obtain the final representation by concatenating the claim representation, report representation, and explanation representation as follows:

$$\mathbf{h}_D = \mathrm{Max}([\mathbf{h}_1; \mathbf{h}_2; \cdots; \mathbf{h}_{|\mathcal{D}|}]) \tag{6}$$

$$\mathbf{h}_E = \mathrm{Max}([\mathbf{h}_1^{''}; \mathbf{h}_2^{''}; ...; \mathbf{h}_K^{''}]) \tag{7}$$

$$\mathbf{h}^\dagger = [\mathbf{h}_c; \mathbf{h}_D; \mathbf{h}_E] \tag{8}$$

where $\mathbf{h}_D$ denotes the integrated representation of all report sentences, $\mathbf{h}_E$ denotes the integrated representation of all explainable sentences. $\mathbf{h}^\dagger$ denotes the final representation for veracity prediction. K denotes a hyperparameter controlling the maximum number of selected reports. Similar to Eq. (2), $\mathbf{h}_k^{''} = \mathrm{Max}([\mathbf{h}_{k,1}^{''}; \mathbf{h}_{k,2}^{''}; \cdots; \mathbf{h}_{k,|d_k^{'}|}^{''}])$ is the k-th report representation in the extracted explanations.

Finally, $\mathbf{h}^\dagger$ is fed into a multi-layer perceptron (MLP) layer to predict the veracity label as follows:

$$\hat{y} = \mathrm{softmax}(\mathrm{MLP}(\mathbf{h}^\dagger)) \tag{9}$$

4.5 Model Training

It is inefficient to train report selection, explainable sentence extraction, and veracity prediction independently, considering their implicit correlations and the pipeline for explainable fake news detection in the real world (Kotonya and Toni, 2020a). Thus, we jointly optimize these three sub-tasks in an end-to-end model. For model training, we minimize the overall loss $\mathcal{L}_{all}$ as follows:

$$\mathcal{L}_D = -\sum_i y_i^d \log(\hat{y}_i^d) \tag{10}$$

$$\mathcal{L}_S = -\sum_k \sum_t y_{k,t}^s \log(\hat{y}_{k,t}^s) \tag{11}$$

$$\mathcal{L}_C = -y\log(\hat{y}) \tag{12}$$

$$\mathcal{L}_{all} = \beta_D \mathcal{L}_D + \beta_S \mathcal{L}_S + \beta_C \mathcal{L}_C \tag{13}$$

where $\mathcal{L}_D$, $\mathcal{L}_S$, and $\mathcal{L}_C$ denote the cross-entropy loss for check-worthy report selection, explanation generation and veracity prediction tasks, respectively. y_i^d and $\hat{y}_i^d$ denote the gold and predicted label of reports, respectively. $y_{k,t}^s$, and $\hat{y}_{k,t}^s$ denote the ground truth and the predicted probability of the sentence for explanation, respectively. y and $\hat{y}$ denote the ground truth and predicted veracity probability of the claim, respectively. β denotes the trade-off parameter, controlling the task importance in our work. We can automatically assign β_D, β_S, and β_C with proper values using the adaptive strategy, rather than the grid search (see Appendix B).

5 Experiments

5.1 Datasets and Settings

To the best of our knowledge, there is no public dataset on raw reports available for this task. Thus, we collect two explainable datasets, i.e., RAWFC and LIAR-RAW, referring to two different fact-checking sites (i.e., Snopes[3] and Politifact[4]) for gold labels, respectively. For RAWFC, we constructed it from scratch by collecting the claims from Snopes and *relevant raw reports* by retrieving claim keywords. For LIAR-RAW, we extended the public dataset LIAR-PLUS (Alhindi et al., 2018) with *relevant raw reports*, containing *fine-grained* claims from Politifact. We process and separate these datasets into train/valid/test sets by 8:1:1 following the same setting in (Atanasova et al., 2020). More details are illustrated in Appendix A.

Dataset	RAWFC	LIAR-RAW
Claim	2,012	12,590
# pants-fire	-	1,013
# false	646	2,466
# barely-true	-	2,057
# half-true †	671	2,594
# mostly-true	-	2,439
# true	695	2,021
Veracity Label	3	6
Explain sentence		
# min	1	1
# max	110	209
# avg	18.4	4.1
Report per claim		
# min	1	1
# max	30	30
# avg	21.0	12.3
Sentence per report		
# min	1	1
# max	155	59
# avg	7.4	5.5

Table 1: Statistics of datasets. # half-true † is also denoted as # half in RAWFC. The number of oracles in datasets isn't pre-defined.

For experimental setup, we initialized word embeddings with the base uncased DistilBERT (Sanh et al., 2019) and $d = 768$ dimensions. The hidden size of LSTM is set to 384. We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-5 and the mini-batch size is set to 1 to minimize joint cross-entropy loss. The maximum number K of selected reports for each claim is empirically set to 12 and 18 for RAWFC and LIAR-RAW, respectively. We use a soft threshold $\varepsilon_i = 1/|d_i'|$ for selection while empirically setting the maximum number of oracle sentences to 30 and 55 for RAWFC and LIAR-RAW, respectively. We set the dropout rate to 0.4 before final prediction and the maximum number of training epochs to 8. For evaluation, we employ macro-averaged precision (P), recall (R), and F1 score (macF1) for veracity prediction, and use ROUGE-N F1 score ($N \in \{1, 2, L\}$) and the human evaluation to evaluate the quality of explanations. *Note that fact-checked reports are not required during inference in our model.*

5.2 Veracity Prediction Performance

Table 2 compares veracity prediction results with the following strong baselines: 1) **SVM** (Pedregosa et al., 2011): This uses bag-of-words features to train SVM-based model for fake news detection; 2) **CNN** (Wang, 2017): This incorporates available metadata features to enhance representation learning; 3) **RNN** (Rashkin et al., 2017): This learns representation from word sequences without external resources; 4) **DeClarE** (Popat et al., 2018): This combines word embeddings from the claim,

[3] www.snopes.com

[4] www.politifact.com

Model	RAWFC			LIAR-RAW		
	P(%)	R(%)	macF1(%)	P(%)	R(%)	macF1(%)
SVM (Pedregosa et al., 2011)	32.33	32.51	31.71	15.78	15.92	15.34
CNN (Wang, 2017)	38.80	38.50	38.59	22.58	22.39	21.36
RNN (Rashkin et al., 2017)	41.35	42.09	40.39	24.36	21.20	20.79
DeClarE (Popat et al., 2018)	43.39	43.52	42.18	22.86	20.55	18.43
dEFEND (Shu et al., 2019)	44.93	43.26	44.07	23.09	18.56	17.51
SentHAN (Ma et al., 2019)	45.66	45.54	44.25	22.64	19.96	18.46
SBERT-FC (Kotonya and Toni, 2020b)	51.06	45.92	45.51	24.09	22.07	22.19
GenFE (Atanasova et al., 2020)	44.29	44.74	44.43	28.01	26.16	26.49
GenFE-MT (Atanasova et al., 2020)	45.64	45.27	45.08	18.55	19.90	15.15
CofCED	**52.99**	**50.99**	**51.07**	**29.48**	**29.55**	**28.93**

Table 2: Experimental results of veracity prediction merely using raw reports ($p < 0.05$ under t-test).

report, and source to access the credibility of the claim; 5) **dEFEND** (Shu et al., 2019): This utilizes GRU-based model for veracity prediction with explanations; 6) **SentHAN** (Ma et al., 2019): This represents each sentence based on sentence-level coherence and semantic conflicts with the claim; 7) **SBERT-FC** (Kotonya and Toni, 2020b): This uses SentenceBERT (SBERT) for encoding and detects fake news based on the top-K ranked sentences; 8) **GenFE/GenFE-MT** (Atanasova et al., 2020): This detects fake news independently or jointly with explanations in the multi-task set-up.

Table 2 demonstrates the detection performance of our proposed CofCED compared with existing strong baselines in terms of precision, recall and macro F1 (macF1). From this table, we can observe that CNN and RNN outperform SVM on both datasets, indicating that deep learning methods can better capture semantic and syntactic features from raw reports. By attentively aggregating multiple features from the claim, reports, and source to estimate the veracity, dEFEND, DeClarE and SentHAN achieve better performance on RAWFC but slightly worse results on LIAR-RAW, because fine-grained labels contained in LIAR-RAW make it more challenging.

SBERT-FC and GenFE outperform SentHAN and dEFEND on both datasets, demonstrating the superiority of pre-trained models. GenFE-MT performs better than GenFE on RAWFC, but much worse than other baselines on LIAR-RAW, implying the challenge of fine-grained fake news detection with explanation generation in the multi-task setting. Generally, CofCED consistently achieves much better performance on RAWFC and LIAR-RAW, demonstrating the superiority of CofCED in combining report selection, explainable sentence extraction and veracity prediction for fake news detection directly on raw reports, alleviating the dependency on fact-checked reports.

5.3 Ablation Study

To evaluate the impact of each component, we conduct ablation experiments for CofCED by removing the following key components: 1) **RS** denotes report selection; 2) **SE** denotes sentence selection; 3) **RS&SE** denotes RS and SE; 4) Four semantic features: **claim relevance**, **richness**, **salience**, and **non-redundancy**, for sentence selection.

As shown in Table 3, CofCED significantly outperforms CofCED w/o $*$ ($*$ indicates a component) on both datasets, demonstrating all components contribute to the effectiveness of CofCED in detecting fake news. Specifically, CofCED's performance significantly decreases without RS&SE because there is noise in raw reports, affecting the veracity prediction. CofCED w/o SE performs much worse than the others because irrelevant or redundant information contained in such reports may weaken the effect of evidence for detection; CofCED w/o RS also achieves worse performance than CofCED because noisy reports may affect sentence selection and model training. Furthermore, the performance of CofCED w/o claim relevance significantly decreases, highlighting the importance of selecting claim-relevant evidence for final prediction. CofCED outperforms CofCED without these four features for sentence selection, respectively, demonstrating they contribute to extracting explainable sentences for fake news detection from different perspectives.

5.4 Explanation Evaluation

Table 4 reports the ROUGE results of the extracted explanations regarding word overlapping. The ROUGE F1 score is employed to evaluate their qualities comparing with the following strong baselines: 1) **LEAD-N** (Nallapati et al., 2017): This uses the first N sentences as explanation and $N = 5$; 2) **Oracle** (Atanasova et al., 2020): This typically presents the best greedy approximation of the gold

Model	RAWFC			LIAR-RAW		
	P(%)	R(%)	macF1(%)	P(%)	R(%)	macF1(%)
CofCED w/o RS&SE	45.01	45.02	44.98	25.69	24.55	24.80
CofCED w/o SE	52.27	46.36	43.80	27.59	23.81	23.74
CofCED w/o RS	49.26	46.92	46.37	27.08	25.32	25.52
CofCED w/o non-redundancy	48.80	46.98	47.48	26.54	27.36	26.65
CofCED w/o salience	43.96	49.24	46.44	26.36	24.88	25.23
CofCED w/o richness	48.08	47.50	47.12	27.06	25.82	26.05
CofCED w/o claim relevance	45.66	45.25	45.28	26.42	24.01	24.88
CofCED	**52.99**	**50.99**	**51.07**	**29.48**	**29.55**	**28.93**

Table 3: Ablation study results of our veracity prediction on test sets; w/o denotes 'without'.

Model	RAWFC			LIAR-RAW		
	ROU-1	ROU-2	ROU-L	ROU-1	ROU-2	ROU-L
LEAD-N	19.52	4.54	17.26	9.84	0.40	7.20
Oracle	37.62	13.22	34.67	25.50	9.28	22.61
EXTABS (Kotonya and Toni, 2020b)	-	-	-	18.85	3.61	12.90
dEFEND (Shu et al., 2019)	19.95	5.08	17.21	17.03	3.26	11.42
GenFE-MT (Atanasova et al., 2020)	18.23	7.12	17.32	**23.08**	3.67	12.10
CofCED w/o non-redundancy	27.32	9.06	23.19	17.96	3.54	12.43
CofCED w/o salience	26.67	7.44	21.02	17.27	3.41	11.69
CofCED w/o richness	25.75	8.66	21.87	17.23	3.44	12.10
CofCED w/o claim relevance	25.56	8.07	20.73	17.08	3.31	11.25
CofCED w/o RS	26.64	8.96	22.69	17.51	**3.72**	**13.20**
CofCED	**27.62**	**9.32**	**23.57**	17.14	3.49	12.96

Table 4: ROUGE results of the generated explanation. ROU-N ($N \in \{1, 2, L\}$) denotes the ROUGE-N F1 score that evaluates the token overlap between the explanation and human justifications. RAWFC is not suitable for EXTABS because its gold justification is too long to train an abstractive-summarization model.

explanation with sentences extracted from reports; 3) **EXTABS** (Kotonya and Toni, 2020b): This uses extractive-abstractive summarization model pre-trained on extra news articles and summaries dataset before fine-tuning (Liu and Lapata, 2019); 4) **dEFEND**: This uses internal attention weights for explanations; 5) **GenFE-MT**: This incorporates explanation generation using pre-trained models.

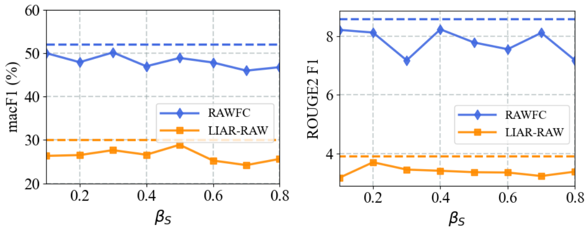
Overall, CofCED achieves the state-of-the-art performance on RAWFC and comparable ROUGE scores with GenFE-MT on LIAR-RAW, suggesting that our CofCED can effectively distill explainable sentences that contributes to the final veracity prediction, as shown in Table 2. Specifically, the ROUGE results of LEAD-N and Oracle on RAWFC and LIAR-RAW indicate that generating explanations for fine-grained fake news detection is a more complex challenge. EXTABS obtains competitive results on LIAR-RAW due to additional news and summaries datasets for abstractive summarization but it cannot deal with long justifications. GenFE-MT performs much better than dEFEND on both datasets, indicating the advantage of pre-trained models in generating explanation from raw reports but failing to trade off both tasks regarding Table 2. For ablation results, we observe that some ablations of CofCED achieve slightly better ROUGE scores but much worse veracity predictions on LIAR-RAW, indicating these four features can effectively select explainable sentences to enhance fake news detection. Besides, CofCED performs better on RAWFC while only comparable on LIAR-RAW than CofCED w/o RS, implying that generating explanations for fine-grained veracity labels is much more challenging regarding word overlapping. We further conduct human evaluations as shown in Appendix D. In summary, our CofCED can effectively generate accurate explanations from raw reports and all components contribute to focusing on veracity prediction.

5.5 Case Study

For in-depth analysis, we further explore the process of CofCED in selecting explainable sentences. We normalized scores for each abstract feature, obtaining its overall probability for explaining detection results. As shown in Table 5, given a false claim about COVID-19, the top two sentences with higher overall scores refute the claim from different perspectives and the last two sentences with 0.3 and 0.2 overall scores contribute less to the veracity prediction. The separated terms, i.e., claim relevance, richness, salience, and non-redundancy, in Eq. (4) are clearly visualized for seeking the major factor responsible for the classification of each sentence. In addition to being a state-of-the-art method

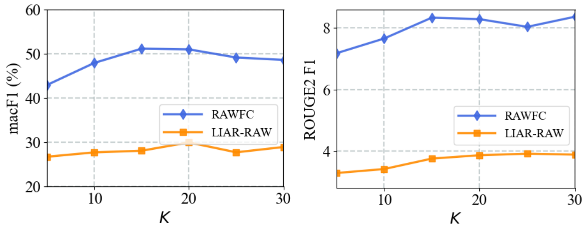
Claim: *Dr. Tasuku Honjo said that COVID-19 was "man-made" at a lab in Wuhan, China.* [**Prediction:** False] **Explanation:** Honjo did not work at the Wuhan Institute of Virology, he did not say that COVID-19 was "invented" or "man-made," and the Twitter account posting similar claims does not belong to the Nobel Prize winner. In addition, this rumor is all based on the unfounded notion that COVID-19 was created as a bioweapon. (...)	Relevance	Richness	Salience	Non-redant	**Overall**	
[1] TOKYO, May 6 (Xinhua) – Japanese Nobel laureate Tasuku Honjo have refuted claim that China manufacture the novel coronavirus, say those rumor be "dangerously distract."	0.9	0.6	0.8	0.9	**0.9**	√
[2] Actually, the professor don't have a Twitter account.	0.7	0.5	0.6	0.9	**0.6**	√
[3] The 2018 Nobel laureate encourage Japanese authority to adopt a more proactive approach.	0.3	0.5	0.4	0.8	**0.3**	×
[4] China will have a big role to play. ...	0.2	0.2	0.1	0.7	**0.2**	×

Table 5: Our visualization of explanation extraction from raw reports. Each row is a sentence in raw reports. The score in the columns are normalized from each of the abstract features in Eq. (4), and the last column is the final probability explaining to detection results.



(a) Veracity prediction (b) Explanation generation

Figure 3: Results of CofCED under different values of the trade-off parameter β_S and $\beta_C = 1 - \beta_S$. The colored dashed horizontal lines denote the performance of CofCED with our adaptive weighting.



(a) Veracity prediction (b) Explanation generation

Figure 4: Results of CofCED under different values of the maximum number K for report selection.

for explainable fake news detection, CofCED has the additional superiority of being very explainable for sentence extraction. Thus, such visualization increases the transparency of the system and the credibility of generated explanations for verdicts.

5.6 Parameter Sensitivity Study

We further investigate the impact of the trade-off parameter β in Eq. (13) on CofCED using the grid search. For brevity, Fig. 3 only presents the results for a) veracity prediction and b) explanation generation on development sets when β_S varies and $\beta_D = 0.5$ is temporarily fixed. We also tried various $\beta_D \in [0.1, 0.8]$ and consistently achieved similar results. By varying the value of β_S from 0.1 to 0.8, our model achieves better performances on one task but poorer results on the other. This is because these tasks show different importance and priority for the final performance over time. By contrast, our CofCED with our proposed multi-task adaptive weighting (MAW) (i.e., the colored dashed horizontal lines) consistently achieves better performance. Thus, these results demonstrate that CofCED with MAW can effectively find better weights for explanation generation and veracity prediction in multi-task learning, alleviating the labor for the grid search for trade-off parameters.

To examine the impact of the maximum number of selected reports on CofCED, we conduct experiments by varying K while fixing other hyper-parameters on the development sets of RAWFC and LIAR-RAW. As shown in Fig. 4, we can see that too few raw reports generally cause performance reduction because the noise in the raw reports may impose the model training bias. Since too many raw reports will cause the out of memory problem, we empirically choose a proper value in this study, i.e., K is set to 12 and 18 for RAWFC and LIAR-RAW, respectively. Note that ε is a soft threshold that can be automatically assigned regarding the total number of report sentences.

6 Conclusion

We present a coarse-to-fine cascaded evidence-distillation (CofCED) neural network for explainable fake news detection that achieves the best detection performance and distills accurate veracity explanations directly from raw reports. Besides, CofCED has the additional advantage of being explainable in producing veracity explanations, explicitly considering the semantic features, e.g., claim relevance, richness, salience, and non-redundancy. Experimental results on real-world datasets demonstrate the effectiveness of CofCED for explainable fake news detection utilizing the wisdom of crowds, effectively mitigating the dependency on fact-checked reports.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China through grants No.61976102, No.U19A2065, and No.61902145. This work is partly supported by the International Cooperation Project (20220402009GH) and Science & Technology Development Program (20210508060RQ), Jilin Province. This work is partly supported by HKBU One-off Tier 2 Start-up Grant (RCOFSGT2/20-21/SCI/004), Hong Kong RGC ECS (22200722).

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: improving fact-checking by justification modeling. In *FEVER*, pages 85–90.

Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *ACL*, pages 7352–7364.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP-IJCNLP*, pages 4685–4697.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and et al. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803. PMLR.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proc. AAAI Conf. Artif. Intell.*, volume 35, pages 81–89.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In *ICCL*, pages 1546–1557.

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

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *ICCL*, pages 5430–5443.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *EMNLP*, pages 7740–7754.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *CVPR*, pages 1871–1880.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3730–3740.

Yunfei Long. 2017. Fake news detection through multi-perspective speaker profiles. In *ACL*.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *ACL*, pages 505–514.

Jing Ma, Wei Gao, Shafiq Joty, and et al. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *ACL*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *WWW*, pages 585–593.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*, pages 6859–6866.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *WWW*, pages 1003–1012.

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

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv:1910.01108*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *SIGKDD*, pages 395–405.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pages 809–819.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL (short)*, pages 422–426.

Lianwei Wu, Yuan Rao, Ling Sun, and Wangbo He. 2021. Evidence inference networks for interpretable claim verification. In *AAAI*, pages 14058–14066.

Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *WWW*, pages 3600–3604.

Appendices

A Dataset Details

Existing benchmarks for explainable fake news detection collected official debunked reports written by journalists as evidence for fake news detection (Kotonya and Toni, 2020a), which is labor-intensive and relatively inefficient. However, debunked reports are not always available for breaking news and are mixed up with raw reports, which may contain more semantically irrelevant and redundant information. To the best of our knowledge, there is no available explainable dataset based on crowds of raw reports to detect fake news before official reports published. Thus, existing datasets are not suitable for most real-life scenarios, especially when the fact-checked reports are not always available. To address this issue, we collect two new datasets, i.e., RAWFC and LIAR-RAW, considering a more general situation of detecting and explaining fake news with relevant raw reports.

Note that we construct RAWFC and LIAR-RAW with gold labels referring to Snopes[5] and Politifact[6], respectively. RAWFC is constructed from scratch as follows and LIAR-RAW are extended with raw reports based on LIAR-PLUS (Alhindi et al., 2018). Besides, we pre-processed LIAR-RAW similar to RAWFC. The detailed statistics of datasets are shown in Table 1.

A.1 Data Collection and Processing.

We crawled claims with their veracity labels and relevant fact-checked reports that can be regarded as gold explanations from Snopes. For each claim, we extracted the claim-related keywords as the search query and used Google API to retrieve the top 30 relevant raw reports. To mitigate the dependency on fact-checked reports, we filtered out reports from fact-checking sites and removed the raw reports published after the publication time of the fact-checked report. We further removed the summary from the remaining articles and improved the quality of the dataset with data cleanings, e.g., removing reports containing less than 5 words or more than 3000 words. Finally, we standardized the original labels for 3-way classification: {*true*, *false*, *half*}, i.e., {true, correct attribute, mostly true} → *true*, { false, misattribute, mostly false } → *false*, {mixture, unproven } → *half*. Each sentence is annotated as evidence or not according to their similarities with the gold explanation, where we greedily extract sentences that achieve the high cosine similarity and ROUGE F1 score, referred to as *oracles*.

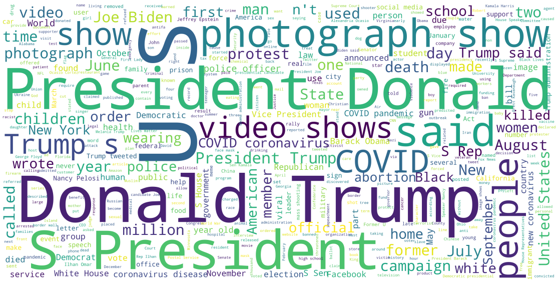


Figure A.1: The word cloud of our RAWFC.

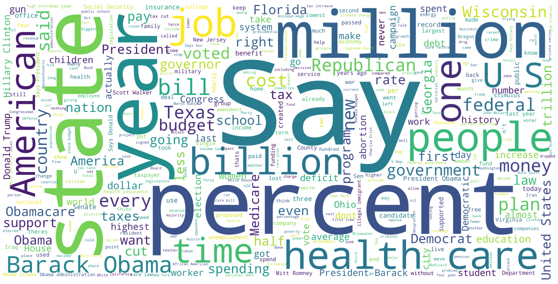


Figure A.2: The word cloud of our LIAR-RAW.

A.2 Evidential Sentence Annotation.

To help produce explanations from external raw reports, each sentence in the article is annotated as evidence or not. Different from selecting evidential sentences based merely on the ROUGE score (Lin, 2004) with gold explanations (Atanasova et al., 2020), we propose a more practical approach to annotate sentences according to both textual-level and semantic-level similarities.

For each candidate sentence, we adopt two metrics to assess whether it should be selected or not: 1) *ROUGE* measures the textual-level similarity regarding the gold explanation in terms of the n-gram overlap; and 2) *Cosine* measures the semantic similarity regarding the gold explanation. Formally, for a candidate sentence $s_{i,j} \in d_i = \{s_{i,j}\}_{j=1}^{|d_i|}$ and its corresponding explanation sentences set $E = \{e_1, e_2, ..., e_n\}$, we define the n-gram overlap function $f^{ROU}(s_{i,j}, e_i)$ and semantic similarity $f^{COS}(s_{i,j}, e_i)$ as follows:

$$f^{ROU}(s_{i,j}, e_i) = \frac{|n\text{-grams}(s_{i,j}) \ \cap \ n\text{-grams}(e_i)|}{|n\text{-grams}(e_i)|} \tag{A.1}$$

$$f^{COS}(s_{i,j}, e_i) = \cos(h_{s_{i,j}}, h_{e_i}), \tag{A.2}$$

[5] www.snopes.com

[6] www.politifact.com

Standardized Label	Train	Valid	Test
true	561	67	67
false	514	66	66
half	537	67	67

Table A.1: Label statistics of claims in RAWFC.

Fine-grained Label	Train	Valid	Test
pants-fire	812	115	86
false	1,958	259	249
barely-true	1,611	236	210
half-true	2,087	244	263
mostly-true	1,950	251	238
true	1,647	169	205

Table A.2: Label statistics of claims in LIAR-RAW.

where $h_{s_{i,j}}$ and h_{e_i} is the sentence representation encoded by SBERT (Reimers and Gurevych, 2019). We calculate the textual similarity in terms of *ROUGE-1*, *ROUGE-2*, and *ROUGE-L* F1 scores, respectively; we also calculate the semantic similarity in terms of *Cosine*. For sentence labeling, we empirically set the thresholds of *ROUGE-1*, *ROUGE-2*, and *ROUGE-L* F1 scores to 0.1, 0.0, and 0.1, respectively, and the threshold of *Cosine* to 0.6. Finally, we accepted the sentences that exceed all given thresholds as gold explanation sentences, i.e., *oracle*. The label statistics of claims in RAWFC and LIAR-RAW are displayed in Table 1 and Table A.2, respectively. Moreover, we also visualized their word clouds, as shown in Fig. A.1 and Fig. A.2, respectively.

B Multi-task Adaptive Weighting

Inspired by prior work (Chen et al., 2018; Liu et al., 2019), we further propose a simple yet effective remedy, namely Multi-task Adaptive Weighting (MAW), to automatically keep a dynamic balance among tasks for different benchmark datasets. We define the weighting function $\beta_k(t)$ as follows:

$$\beta_k(t) = \frac{N_k \exp[f_k(t)g(t)]}{\sum_i \exp[f_i(t)g(t)]} \quad \text{(B.1)}$$

$$f_k(t) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}, g(t) = \frac{\log(t-2)}{T} \quad \text{(B.2)}$$

where $\beta_k = \beta_k(t), k \in \{D, S, C\}$ and $f_k(t)$ represents the loss rate for task where t is an iteration step; $g(t)$ is a global function that can generate a growth value, contributing to an optimal balance between tasks, since a large T can result in a more even distribution between different tasks. $T = 8$ denotes an initial temperature to control the softness of task weighting similar to (Caruana, 1997). $N_k = 3$ indicates the total number of sub-tasks. We simply initialize $\beta_k = 0.5$ and update the average loss over each iteration.

Algorithm 1: CofCED

```
Input: A set of training instances $\{(c, \mathcal{D})\}$;
       Maximum selection number $K$;
       Thresholds $\varepsilon$.
Output: Veracity label $\hat{y}$; Check-worthy
        report labels $\hat{Y}^d$; Explainable
        sentence labels $\hat{Y}^s$; Generated
        Explanation $\hat{E}$
Initialize $\beta_D = \beta_S = \beta_C = 0.5$, if $t \leq 2$;
for each instance $(c, \{\{s_{i,j}\}_{j=1}^{|d_i|}\}_{i=1}^{|\mathcal{D}|})$ do
    {Hierarchical Encoding}
    $\mathbf{h}_c, \mathbf{h}_{i,j} \leftarrow$ DistilBERT;
    $\mathbf{h}_i \leftarrow$ Eq. (2);
    {Task 1: Report Selection}
    $\hat{y}_i^d, \{d'_k\}_{k=1}^K \leftarrow K$; Eq. (3);
    $\mathbf{h}_D = \text{Max}([\mathbf{h}_1; \mathbf{h}_2; ...; \mathbf{h}_{|\mathcal{D}|}])$
    {Task 2: Explainable Sentence
     Extraction}
    for each report $d_k$ in $\{d'_k\}_{k=1}^K$ do
        $\hat{y}_{k,t}^s \leftarrow$ Eq. (4);
        $\{s_{k,t}\}_{t=1}^{|d'_k|}, \{\mathbf{h}''_{k,t}\}_{t=1}^{|d'_k|} \leftarrow \hat{y}_{k,t}^s > \varepsilon_k$
    Explanations: $\hat{E} = \{\{s_{k,t}\}_{t=1}^{|d'_k|}\}_{k=1}^K$,
    $\mathbf{h}''_k = \text{Max}([\mathbf{h}''_{k,1}; \mathbf{h}''_{k,2}; \cdots; \mathbf{h}''_{k,|d'_k|}])$;
    $\mathbf{h}_E = \text{Max}([\mathbf{h}''_1; \mathbf{h}''_2; ...; \mathbf{h}''_K])$;
    {Task 3: Veracity Prediction}
    $\mathbf{h}^\dagger = [\mathbf{h}_c; \mathbf{h}_D; \mathbf{h}_E]$
    Verdicts: $\hat{y} \leftarrow$ Eq. (9);
{Multi-task Training}
Optimize
  $\mathcal{L}_{all} = \beta_D \mathcal{L}_D + \beta_S \mathcal{L}_S + \beta_C \mathcal{L}_C \leftarrow$ Eq.
  (10,11,12);
Update $\beta_D, \beta_S, \beta_C$;
```

C CofCED Algorithm

Algorithm 1 shows our training procedure.

D Human Evaluation for Explanations

We also study the explanation quality by human evaluation referring to (Atanasova et al., 2020). Provided with three types of explanations, i.e., human justification, veracity explanation generated by CofCED, and the ones generated by GenFE-MT,

RAWFC			
Annotator	Gold	Exp-GenFE-MT	Exp-CofCED
<Informativeness>			
# 1	**1.38**	2.17	1.89
# 2	**1.63**	2.32	2.01
# 3	**1.24**	1.76	2.05
ALL	**1.42**	2.08	1.98
<Readability>			
# 1	**1.74**	1.98	1.81
# 2	**1.15**	1.76	1.63
# 3	**1.97**	2.35	2.07
ALL	**1.62**	2.03	1.84
<Overall>			
# 1	**1.54**	1.98	2.13
# 2	**1.43**	1.76	1.73
# 3	**1.60**	2.24	1.91
ALL	**1.52**	1.99	1.94
LIAR-RAW			
<Informativeness>			
# 1	**1.27**	1.91	1.82
# 2	**1.55**	2.09	1.63
# 3	**1.12**	1.72	1.46
ALL	**1.31**	1.91	1.64
<Readability>			
# 1	**1.13**	2.29	1.78
# 2	**1.38**	2.25	2.12
# 3	**1.24**	1.94	2.02
ALL	**1.25**	2.16	1.97
<Overall>			
# 1	**1.33**	1.96	1.68
# 2	**1.49**	2.12	1.94
# 3	**1.51**	2.35	2.08
ALL	**1.44**	2.14	1.90

Table C.1: Mean Average Ranks (MAR) of the explanations for each three evaluation criteria on RAWFC and LIAR-RAW, respectively. Gold denotes the explanations come from the justification, Exp-GenFE-MT denotes the explanations generated by GenFE-MT, and Exp-CofCED denotes the explanations generated by our CofCED. Best performances are shown in bold, and the second ones are underlined.

three English-speaking adult annotators were asked to rank them with 1–Good, 2–Medium, 3–Poor, according to three different criteria. To keep clear and simple, we use the following criteria:

- **Informativeness**. The explanation contains much evidential information that contributes to fake news detection.

- **Readability**. The explanation is easy to understand.

- **Overall**. The explanation is ranked based on their overall quality.

Dataset	P(%)	R(%)	macF1(%)
RAWFC	84.28	79.29	81.71
LIAR-RAW	14.98	61.06	24.06

Table E.1: Our results on report classification.

For the annotation settings, we randomly sample a set of 40 instances from the test set and prepare three candidate explanations without any other information about these explanations. All of annotators work independently.

Table C.1 shows the mean average results from the manual evaluation. We also compute Krippendorff's inter-annotator agreement (Atanasova et al., 2020) and obtain 0.37 for Informativeness, 0.43 for Readability, 0.31 for Overall. From the results, we can see that the human justification (Gold) achieves the best quality and our Exp-CofCED achieves better quality of explanations than Exp-GenFE-MT. These results suggest that the ROUGE results in Table 4 may be not sufficient for evaluating veracity explanations because the ROUGE score only accounts for word overlapping. Besides, the performance of veracity prediction in Table 2 also verifies the effectiveness of explanations in improving fake news detection. In summary, our proposed CofCED can significantly improve final fake news detection with overall better veracity explanations.

E Further Discussion

Table E.1 shows internal results about report classifications regarding precision, recall, and macro F1 score. Our model outperforms better on RAWFC than on LIAR-RAW, indicating that report classification for fine-grained claims is much challenging and further improving this part may contribute to explainable fake news detection. Similarly, Table E.2 shows internal results about explainable sentence classifications. Overall, our CofCED significantly outperforms GenFE-MT but only achieves comparable results on LIAR-RAW in terms of ROUGE scores (Table 4). This is probably because ROUGE scores w.r.t. word overlapping are not sufficient for evaluating the qualities of generated explanations. Thus, we further introduce human evaluation as a complementary measure.

F Example

Examples from RAWFC are shown in Table F.1.

Model	RAWFC			LIAR-RAW		
	P(%)	R(%)	macF1(%)	P(%)	R(%)	macF1(%)
GenFE-MT (Atanasova et al., 2020)	50.62	36.03	42.09	**43.83**	4.27	7.79
CofCED	**55.56**	**41.67**	**47.62**	14.29	**22.22**	**17.39**

Table E.2: Experimental results of explainable sentence classification regarding oracle sentences.

[**Label:** False] **Claim:** *U.S. Rep. Alexandria Ocasio-Cortez started “chain migration" deportation proceedings against First Lady Melania Trump and her parents.*
Explanation: Illegal immigration remained a top issue for U.S. President Donald Trump and continued to divide Americans in mid-2019, all the more so after Trump told several Democratic members of Congress of immigrant parentage, all but one of them born in the United States, they should “go back and help fix the totally broken and crime infested places from which they came.” (...) This is simply not true. For context, “chain migration” is a term used to describe immigration procedures that allow adult U.S. citizens to obtain citizenship for foreign-born adult relatives. Reportedly, the first lady’s parents secured their citizenship through just such a procedure — though we needn’t belabor the point, because everything else in the story is fictional (Melania Trump’s parents aren’t named “Oedipus and Jezebel Beelzebub.”
Raw Report Domain: *www.newsweek.com*
Content: The president have also be criticize for want to end “chain migration”, a program that let U.S. citizen to sponsor immediate family member for legal residency, despite it be the program that Melania Trump use to put her parent Viktor and Amalija Knavs on a path to American citizenship. (...)
Raw Report Domain: *www.washingtonpost.com*
Content: Melania Trump’ s parent be legal permanent resident, raise question about whether they rely on “chain migration” She enjoy put her personal mark on the historic home and have redesign the family live quarter. (...)
Raw Report Domain: *www.kbzk.com*
Content: Melania Trump’s parent, Viktor and Amalija Knavs, also go through the immigration process, use the perjoratively call “chain migration” route the President have criticize. (...) A source with direct knowledge of Melania Trump’s parent and their immigration status previously tell CNN that she have sponsor her parent for their green card, a status that allow them to live and work in the US indefinitely and pave the way for citizenship. (...)

[**Label:** True] **Claim:** *The snakehead fish can survive on land.*
Explanation: On Oct.10, 2019, many readers came across news stories about an invasive species of fish called the snakehead fish that had been discovered in Georgia. While these stories largely dealt with wildlife officials’ attempts to eradicate the species, what caught the attention of most readers were brief mentions of this fish’s unique ability to survive on land. CNN reported: A snakehead fish that survives on land was discovered in Georgia. Officials want it dead An invasive fish species that can breathe air and survive on land has been found in Georgia for the first time. And officials are warning anyone who comes into contact with the species to kill it immediately. The snakehead fish can truly survive on land. Here’s a video of a snakehead in Thailand as it “walks,” crawls, or wiggles its way back to the water.
Raw Report Domain: *www.cbsnews.com*
Content: Northern snakehead be invasive fish that can breathe air and survive for day on land. Lawrenceville, Georgia — Georgia’s Department of Natural Resources have a message for angler: If you catch a northern snakehead, kill it immediately.
Raw Report Domain: *www.nytimes.com*
Content: Snakeheads can survive in freshwater and be describe a predator that can eat tiny animal, and travel across land, live out of water for several day. There have be no end to the creepy description of the snakehead fish, a slimy, toothy, large-jawed animal that can breathe on land and crawl like a snake, in the decade that it have pop up in freshwater lake, pond and river in the United States. (...)

Table F.1: Examples from RAWFC.