Crowd, Expert & AI: A Human-AI Interactive Approach Towards Natural Language Explanation Based COVID-19 Misinformation Detection

Ziyi Kou[1]*, Lanyu Shang[1]*, Yang Zhang[2], Zhenrui Yue[1], Huimin Zeng[1], Dong Wang[1]

[1]Unversity of Illinois at Urbana-Champaign
[2]University of Notre Dame
{ziyikou2, lshang3, zhenrui3, huiminz3, dwang24}@illinois.edu, yzhang42@nd.edu

Abstract

In this paper, we study an explainable COVID-19 misinformation detection problem where the goal is to accurately identify COVID-19 misleading posts on social media and explain the posts with natural language explanations (NLEs). Our problem is motivated by the limitations of current explainable misinformation detection approaches that cannot provide NLEs for COVID-19 posts due to the lack of sufficient professional COVID-19 knowledge for supervision. To address such a limitation, we develop CEA-COVID, a crowd-expert-AI framework that jointly exploits the common logical reasoning ability of online crowd workers and the professional knowledge of COVID-19 experts to effectively generate NLEs for detecting and explaining COVID-19 misinformation. We evaluate CEA-COVID using two public COVID-19 misinformation datasets on social media. Results demonstrate that CEA-COVID outperforms existing explainable misinformation detection models in terms of both explainability and detection accuracy.

1 Introduction

The proliferation of COVID-19 misleading information on social media has posed a serious threat to the public health and online news ecosystem [Shang *et al.*, 2022a]. For example, nearly 800 people have died in the first quarter of 2020 as they believe that consuming bleach could disinfect the body and kill the COVID-19 virus [Islam *et al.*, 2020]. Therefore, the COVID-19 misinformation detection problem has gained increasing research attention and public interest in promoting Good Health and Well-being (i.e., the United Nations' Sustainable Development Goals (SDGs)). However, it is often difficult for AI algorithms to effectively detect the COVID-19 misinformation and explain the detection results due to the lack of accurate COVID-19 professional knowledge in the training data to optimize the AI models. To address this challenge, we focus on a natural language explanation (NLE) based COVID-19 misinformation detection problem where the goal is to accurately identify COVID-19 misinformation on social media and provide explanations for the identified misinformation in natural language.

Several initial efforts have been made to study the explainable COVID-19 misinformation detection problem in machine learning and human-computer interaction (HCI) communities [Ayoub *et al.*, 2021; Kou *et al.*, 2022; Shang *et al.*, 2022a]. We show three types of explanations in Figure 1 for a misleading COVID-19 post to illustrate the limitations of current approaches. For Figure 1a, Ayoub *et al.* [Ayoub *et al.*, 2021] proposed an attention-based AI framework that identifies misleading COVID-19 posts and extracts specific words from the posts as explanations [Jain and Wallace, 2019]. However, people cannot directly understand the reason of the misinformation simply by reading the extracted words due to the lack of semantic context between the words. Kou *et al.* [Kou *et al.*, 2022] tasked online crowd workers to analyze COVID-19 articles (e.g., fact-checking articles, medical literature) and extracted COVID-19 knowledge triples from the articles as explanations (e.g., "bleach" $\xrightarrow{\text{not cure}}$ "COVID-19" in Figure 1b). However, people may still not be able to understand *why* the bleach is incapable of killing COVID-19 in human's body due to the lack of explicit justification on the knowledge fact (e.g., bleach corrodes human tissues). In contrast, the NLE in Figure 1c is an example of the objective of this paper that can effectively address the above limitations by providing critical details in natural human language to explicitly explain the exact reason why the post is misleading.

Motivated by the above observations, we develop CEA-COVID, a tripartite **C**rowd-**E**xpert-**AI** interactive framework to address the NLE-based COVID-19 misinformation detec-

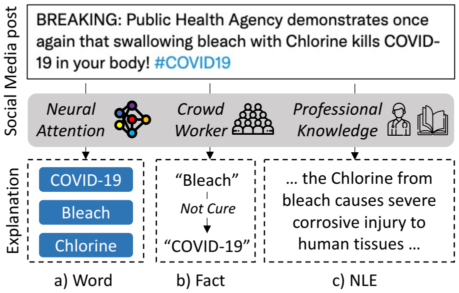

Social Media post
BREAKING: Public Health Agency demonstrates once again that swallowing bleach with Chlorine kills COVID-19 in your body! #COVID19
Neural Attention
Crowd Worker
Professional Knowledge
Explanation
COVID-19
Bleach
Chlorine
"Bleach"
Not Cure
"COVID-19"
... the Chlorine from bleach causes severe corrosive injury to human tissues ...
a) Word
b) Fact
c) NLE

Figure 1: Three Types of COVID-19 Misinformation Explanation

*The first two authors contributed equally to this work.

tion problem. In our solution, we leverage the strengths of three parities: i) crowd workers who often can easily understand COVID-19 posts on social media, ii) COVID-19 expert who have professional COVID-19 knowledge, and iii) AI models that can quickly navigate COVID-19 articles and extract relevant knowledge facts (e.g., the statement in Figure 1c), to collaboratively identify misleading COVID-19 posts and generate NLEs for the identified misinformation. However, the design of such an NLE-based COVID-19 misinformation detection framework is non-trivial due to two research challenges: i) how to design a crowd-AI collaboration model to analyze the COVID-19 posts and extract highly relevant COVID-19 knowledge facts from the COVID-19 articles as the NLE for the post? 2) How to leverage the professional knowledge from the COVID-19 experts to justify the uncertain COVID-19 posts identified by crowd workers and noisy knowledge facts extracted by the AI model?

To address the first challenge, we develop a novel logic-driven COVID-19 knowledge graph that can effectively identify COVID-19 knowledge facts from COVID-19 articles by exploring the semantic logical relations between the COVID-19 posts and the knowledge facts. To address the second challenge, we design an expert-guided knowledge updating strategy to dynamically add the latest COVID-19 knowledge facts to our knowledge graph by leveraging the professional COVID-19 knowledge from the COVID-19 experts. To our best knowledge, CEA-COVID is the first human-AI interactive framework to solve the NLE based COVID-19 misinformation detection problem by jointly leveraging the intelligence from crowd workers, COVID-19 experts and AI models. We evaluate CEA-COVID on two large-scale COVID-19 misinformation datasets on social media. The results show that CEA-COVID not only correctly identifies the misleading COVID-19 posts but also generates more explicit and accurate NLEs for the detected posts than the state-of-the-arts.

2 Related Work

Misinformation Detection and Explanation. Explainability is a critical aspect of misinformation detection, which aims to provide justifications for the misinformation detection results [Brand *et al.*, 2021; Karagiannis *et al.*, 2020; Shang *et al.*, 2022b]. A few explainable misinformation detection solutions have been proposed to explain the identified misinformation with various types of explanations, including top misleading keywords [Ayoub *et al.*, 2021] and knowledge fact triples [Cui *et al.*, 2020]. However, these solutions cannot be applied to solve our problem since they are insufficient to generate NLEs that contain detailed and explicit explanations of the detected misinformation. Moreover, several recent efforts in natural language processing have been developed to generate NLEs for question answering tasks [Danilevsky *et al.*, 2020]. Current NLE generation approaches always require a large number of accurate NLE training labels (e.g., several millions of open-domain questions) that are usually annotated by crowd workers. However, crowd workers are often incapable of generating NLE labels for COVID-19 posts due to their lack of COVID-19 professional knowledge. We develop a crowd-expert-AI interaction framework that leverages the collaboration between crowd workers and COVID-19 experts to jointly generate the NLEs for COVID-19 posts by providing the COVID-19 knowledge facts retrieved from COVID-19 articles by the AI model.

Crowdsourcing Intelligence. Our work is also closely related to leveraging collective human intelligence from a large number of crowd workers to improve AI models. Traditional crowdsourcing solutions often utilize crowdsourcing efforts for conventional annotation tasks, such as object annotation [Barbu *et al.*, 2019] and image captioning [Zhou *et al.*, 2020]. A few recent crowdsourcing solutions have been proposed to solve domain-specific annotation problems by leveraging the human background knowledge from crowd workers [Kou *et al.*, 2022; Méndez Méndez *et al.*, 2019]. However, these solutions often require crowd workers to understand the domain-specific crowdsourcing tasks well by reading various professional articles (e.g., COVID-19 articles), which is largely inefficient as crowd workers are not interested in the reading tasks and may generate unexpected annotation noise. We develop a knowledge-independent crowdsourcing strategy that exploits the logic reasoning ability of crowd workers to efficiently extract the semantic logic information from the COVID-19 posts for the NLE generation.

3 Problem Statement

Definition 1. COVID-19 Post (p)**:** A COVID-19 post is a piece of short text from social media that describes a news topic or statement related to COVID-19 (e.g., the post in Figure 1). In particular, we define a set of N COVID-19 posts as $\mathcal{P} = \{p_1, \ldots, p_N\}$.

Definition 2. COVID-19 Post Label (y)**:** A COVID-19 post p_n is considered as misleading ($y_n = 1$) if it contains partial or entire false information. Otherwise, the COVID-19 post is considered as non-misleading ($y_n = 0$). In this paper, we assume that a COVID-19 post only contains one COVID-19 related claim that is either misleading or non-misleading.

Definition 3. COVID-19 Article (a)**:** An article related to COVID-19, such as a COVID-19 related medical publication (e.g., the New England Journal of Medicine) or a debunking article from credible fact-checking websites (e.g., FactCheck.org). Such articles usually contain COVID-19 related natural language texts that explicitly disprove the misinformation in COVID-19 posts. In particular, we define a set of M COVID-19 articles as $\mathcal{D} = \{a_1, a_2, \ldots, a_M\}$, where $a_m = \{s_{m,1}, \ldots, s_{m,L_s}\}$ is the m^{th} article from $\mathcal{D}$ and $s_{m,l}$ is the l^{th} *COVID-19 statement* (i.e., l^{th} sentence) in a_m. Given a COVID-19 post, a COVID-19 statement is considered as a potential NLE if it can be used to partially or completely judge the truthfulness of the posts.

Definition 4. COVID-19 Expert $(\mathcal{X})$**:** The COVID-19 experts are professionals who have COVID-19 medical or healthcare knowledge to fully understand COVID-19 posts.

Definition 5. Crowd Worker $(\mathcal{C})$**:** The crowd workers are a set of users in online crowdsourcing platforms (e.g., Amazon MTurk) who are not proved to have COVID-19 related professional knowledge.

Definition 6. COVID-19 Natural Language Explanation (NLE) (z): Given a COVID-19 post p_n, a COVID-19 NLE is a statement that explicitly explains why the post is misleading or not (e.g., the texts in Figure 1c). In particular, the NLE of p_n is defined as $z_n = \{z_{n,1}, \dots, z_{n,L_z}\}$ where L_z is the total number of words in z_n.

Definition 7. NLE-based COVID-19 Misinformation Detection Model ($\mathcal{M}$): Given a COVID-19 post p_n, our NLE-based COVID-19 misinformation detection model $\mathcal{M}$ is expected to generate 1) the binary prediction to indicate if p_n is misleading or not; 2) the NLE that explains its prediction.

Using the definitions above, our NLE-based COVID-19 misinformation detection problem is formally defined as:

$$\begin{gathered} z_n = \{z_{n,1}, \dots, z_{n,L_z} | p_n, \mathcal{A}, \mathcal{X}, \mathcal{C}\}, 1 \leq n \leq N \\ \arg\max_{\mathcal{M}^*} \prod_{n=1}^{N} \Pr(\hat{y}_n = y_n, \hat{z}_n = z_n | p_n) \end{gathered} \tag{1}$$

where $\hat{y}_n$ is the misinformation prediction and $\hat{z}_n$ is the generated NLE by $\mathcal{M}$. We expect CEA-COVID to maximize both the misinformation detection accuracy and the explainability.

4 Solution

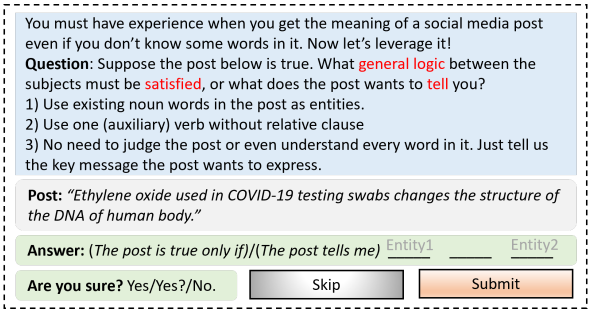
4.1 Crowdsourcing Logical Reasoning Interface

The first module in our scheme is Crowdsourcing Logical Reasoning Interface (CLRI) that is designed to create crowdsourcing tasks for online crowd workers to explain *why* the assigned COVID-19 posts are misleading (or not). We observe that it is challenging for crowd workers to generate reasonable explanations (e.g., the NLE in Figure 1c) even if they are given the truthfulness of the COVID-19 posts due to their lack of professional COVID-19 knowledge.

To solve the above problem, we leverage the general logical reasoning ability of crowd workers to analyze the semantic logic embedded in the COVID-19 posts without requiring them to understand domain-specific terms (e.g., "Chlorine"). We show an example crowdsourcing task in Figure 2. In particular, we hide the binary label of the assigned COVID-19 post and assume all the posts as non-misleading, which effectively saves the time of crowd workers from judging the truthfulness of the post and identifying the misinformation. We formally define the semantic logic below.

Definition 8. Semantic Logic: The semantic logic of a COVID-19 post denotes the relations between different entities in the post regardless of the semantic meaning of the entities. For example, a crowd worker may not know the semantic meaning of "Ethylene oxide" in Figure 2 and hence cannot generate related NLEs. However, it is not difficult for the crowd worker to understand the semantic logic of the post (e.g., "Ethylene oxide somehow damages human DNA").

Given a crowdsourcing task with a COVID-19 post, a crowd worker is expected to generate a piece of semantic logic of the post and submit it as the response. Therefore, the semantic logic of the post is expected to be exactly contradicted or consistent with the truthfulness of the corresponding COVID-19 post (i.e., misleading or non-misleading) (i.e., ground-truth label of the post). We adopt human intelligence from the crowd to create the semantic logic because such logic information is non-trivial to be captured by data-driven AI models that usually lack semantic reasoning ability to generate abstracted logic information [Ayoub *et al.*, 2021]. Given $\mathcal{P}_C = \{p_1, \dots, p_{N_c}\}$ as the set of COVID-19 posts in crowdsourcing tasks, we define $\Omega_c = \{\omega_1, \dots, \omega_{N_c}\}$ as the collected semantic logic from crowd workers.



You must have experience when you get the meaning of a social media post even if you don't know some words in it. Now let's leverage it!
Question: Suppose the post below is true. What general logic between the subjects must be satisfied, or what does the post wants to tell you?
1) Use existing noun words in the post as entities.
2) Use one (auxiliary) verb without relative clause
3) No need to judge the post or even understand every word in it. Just tell us the key message the post wants to express.

Post: *"Ethylene oxide used in COVID-19 testing swabs changes the structure of the DNA of human body."*

Answer: (*The post is true only if*)/(*The post tells me*) Entity1 ______ ______ Entity2 ______

Are you sure? Yes/Yes?/No.

Figure 2: An example crowdsourcing task.

4.2 Logic-driven NLE Graph Network

Given the generated semantic logic of COVID-19 posts from the CLRI, we develop the Logic-driven NLE Graph Network (LNGN) module that extracts high-relevant knowledge facts from the COVID-19 articles as potential NLEs of the posts and leverages the semantic logic as effective regularization. In particular, we first model the set of COVID-19 articles $\mathcal{D}$ as a multi-relational knowledge graph. We extract the pre-annotated entities from COVID-19 posts (e.g., "Chlorine", "human tissues") as initial graph nodes and then identify both the related entities (e.g., compound entities) as graph nodes and their relations as graph edges (e.g., "corrosive injury") by developing heuristic semantic word extraction scripts. The reason of constructing the knowledge graph is to effectively integrate the COVID-19 posts with the structured knowledge facts that are converted from the unstructured human language texts of COVID-19 articles. In particular, given each COVID-19 statement $s_{m,l}$ of the COVID-19 article $a_m \in \mathcal{D}$, we define $V_{m,l} = \{v_1, \dots, v_{K_{m,l}}\}$ and $R_{m,l} = \{r_1, \dots, r_{Q_{m,l}}\}$ as the set of entities and relations extracted from $s_{m,l}$, respectively. If two entities v_{k_1} and v_{k_2} are connected through a relation r_q, we denote them as a *knowledge triple* $T_{m,l,k_1,k_2} = (v_{k_1}, r_q, v_{k_2})$. Therefore, we construct the NLE graph network (NLE-GN) from all COVID-19 articles and denote it as $\mathbb{G} = (\mathcal{V}, \mathcal{R}, \mathcal{T}, \mathcal{A})$ where $\mathcal{V} = \{v_1, \dots, v_K\}$, $\mathcal{R} = \{r_1, \dots, r_Q\}$ and $\mathcal{T} = \{T_1, \dots, T_E\}$ represent the union entities, relations and knowledge triples of $\mathbb{G}$, respectively. $\mathcal{A}$ denotes a $K \times K$ matrix that contains binary values to denote whether two entities from $\mathcal{V}$ are connected or not.

Given the constructed NLE-GN $\mathbb{G}$ and a COVID-19 post p_n, LNGN aims to classify p_n as misleading or not by identifying the relevant knowledge facts with p_n from $\mathbb{G}$. Such knowledge facts can be further traced back to original COVID-19 statements as potential NLE labels. Specifically, LNGN first embeds $\mathbb{G}$ and p_n into high-dimensional embeddings by applying a BERT-based entity encoder to accurately encode each term from $\mathbb{G}$ and p_n into the same vector space. We denote the generated embeddings of the entities, relations, COVID-19 posts and semantic logic as $\widetilde{\mathcal{V}} \in \mathbb{R}^{K\times d}$, $\widetilde{\mathcal{R}} \in \mathbb{R}^{Q\times d}$, $\widetilde{p}_n \in \mathbb{R}^{1\times d}$ and $\widetilde{\omega}_n \in \mathbb{R}^{1\times d}$ respectively.

We then develop a *post-guided knowledge graph propagation strategy* for the LNGN to identify the relevant knowledge facts from $\mathbb{G}$ for the input COVID-19 post p_n by jointly exploring the semantic information of both $\mathbb{G}$ and p_n as $\widetilde{v}_i = \text{ReLU}(\sum_{(j,r,i)\in\mathcal{T}^*} \frac{1}{z_j} W_{i,j}\widetilde{v}_j\mathbb{A}_{i,j})$ where $\widetilde{v}_i$ and $\widetilde{v}_j$ are i^{th} and j^{th} encoded entities from $\widetilde{V}$. $\mathcal{T}^* \subset \mathcal{T}$ denotes the set of knowledge triples consisting of $\widetilde{v}_i$. $\mathbb{A}$ is $K \times K$ matrix derived by jointly aggregating the semantic information from p_n and the adjacent matrix $\mathcal{A}$ with binary values as $\mathbb{A} = (\widetilde{\mathcal{V}}W_1(\widetilde{p}_n)^T + \widetilde{p}_nW_2(\widetilde{\mathcal{V}})^T) \odot \mathcal{A}$. After the knowledge propagation, we generate the embeddings of knowledge triples from $\mathbb{G}$ by multiplying the entity embeddings with the relation embedding as $\widetilde{T}_{i,q,j} = \widetilde{v}_i \odot \widetilde{r}_q \odot \widetilde{v}_j$. Each knowledge triple embedding corresponds to a COVID-19 statement from COVID-19 articles as potential NLE label of p_n. However, it is extremely challenging for the LNGN to accurately identify the relevant knowledge triples due to the lack of direct training supervision on the knowledge triple retrieval process.

To solve the above problem, we leverage the semantic logic ω_n of p_n from the CLRI to effectively retrieve relevant knowledge triples by exploring the logical consistency between ω_n and $\mathcal{T}$. We define the logic consistency as:

Definition 9. Logical Consistency (LC): Given a knowledge triple $T_e \in \mathcal{T}$ and ω_n of p_n, the logical consistency is a binary value (i.e., -1 or 1) to denote if the semantic information of T_e matches with ω_n. For example, the LC is -1 between the example semantic logic in Figure 2 and the knowledge triple ("Ethylene oxide" $\xrightarrow{\text{not affect}}$ "Human DNA").

Given the above definition, the logical consistency between the embedding $\widetilde{T}_e$ and $\widetilde{\omega}_n$ is expected to be high if p_n is non-misleading (i.e., $y_n = 0$). Otherwise, the logical consistency is low because the ω_n is generated by falsely assuming p_n is non-misleading. Therefore, we define the logical consistency loss as $\mathcal{L}_{\text{logic}} = |\cos(\widetilde{\omega}_n, \widetilde{T}_{\text{top}}) - (-1)^{1[y_n=1]}| + \epsilon$ where $\widetilde{T}_{\text{top}} \in \widetilde{\mathcal{T}}$ denotes the embedding with the highest attention score that is generated by transforming the embeddings $\widetilde{\mathcal{T}}$ to 1-dimensional vector as $\text{ATT}_n = \text{Softmax}(W_{\text{att}}(\widetilde{\mathcal{T}})^T)$. We then multiply $\widetilde{\mathcal{T}}$ with ATT_n and apply mean-pooling to generate a single post-level embedding $z_n \in \mathbb{R}^{1\times d}$ that is further transformed to $\hat{y}_n \in \mathbb{R}^2$ as the output probability of p_n being misleading or not. The cross entropy loss $\mathcal{L}_{\text{ce}}$ is finally applied to train LNGN with gradient descent optimization.

4.3 Expert-driven Graph Knowledge Updater

After optimizing the LNGN, the Expert-driven Graph Knowledge Updater (EGKU) aims to update the knowledge facts from $\mathbb{G}$ by efficiently tasking COVID-19 experts to investigate the truthfulness of both the knowledge triples from $\mathbb{G}$ and the semantic logic of COVID-19 posts. However, it is extremely time-consuming if the COVID-19 experts are tasked to check each semantic logic and knowledge triple due to the large amount of available COVID-19 posts and articles. To address the limitation, we only expect COVID-19 experts to investigate specific pairs of semantic logic and knowledge triples by specifying the *logical confidence relation* below.

Definition 10. Logical Confidence Relation (LCR): Given a knowledge triple $T_e \in \mathbb{T}$ and a semantic logic ω_n of p_n, the LCR indicates how confident we believe that they have certain consistent or contradicted local consistency. In particular, the LCR between them is considered to be high if $\cos(\widetilde{\omega}_n, \widetilde{T}_{\text{top}})$ in $\mathcal{L}_{\text{logic}}$ is lower than ξ and higher than $1 - \xi$ where $(0.5 < \xi < 1)$ is the pre-defined threshold. Otherwise, the LCR is considered to be low.

For the pairs of semantic logic and knowledge triples with low LCR, we design a *ranking-based knowledge graph updating strategy* that tasks the COVID-19 experts to judge which semantic information (i.e., semantic logic or knowledge triple) is considered as true. The reason of applying the ranking-based strategy is to minimize the potential opinion bias of COVID-19 experts given the fact that humans usually make more fair estimations if they compare different candidate options rather than solely judging a single option [Burton, 2004]. In particular, if the COVID-19 experts believe the semantic logic contains true COVID-19 related information, the knowledge triple in the pair is removed from NLE-GN while the semantic logic is added. Otherwise, the crowd worker who submitted the semantic logic will be notified to provide more reasonable logic information in the following crowdsourcing tasks. After the COVID-19 experts update the NLE-GN, we levearge the updated NLE-GN to optimize the LNGN again. We repeat such process in LNGN and EGKU for Φ rounds defined as a hyper-parameter of CEA-COVID.

4.4 Knowledge Graph-based NLE Generator

After the above optimization process, we generate the relevant knowledge triples for COVID-19 posts and track the knowledge triples back to corresponding COVID-19 statements. We build the Knowledge Graph-based NLE Generator (KGNG) and optimize KGNG with the retrieved COVID-19 statements as training data. We develop the KGNG based on the multi-head attention-based transformer framework [Ma *et al.*, 2019] and propose a novel *word-level weighted discriminator* as the loss function. In particular, the KGNG takes each COVID-19 post $p_n = \{w_{n,1}, w_{n,2}, \ldots, w_{n,L_p}\}$ as input and outputs the predicted NLE as $\hat{z}_n = \{\hat{z}_{n,1}, \hat{z}_{n,2}, \ldots, \hat{z}_{n,L_z}\}$. Given Top-$K$ retrieved COVID-19 statements from the NLE-GN $S^*_n = \{s_1, s_2, \ldots, s_K\}$, we calculate the weighted per-word cross entropy loss as $\mathcal{L}_{\text{NLE}} = \frac{1}{Z}\sum_w N(w) \times \text{CrossEntropy}(w, \hat{w}) + \mathcal{L}_D(\hat{z}_n, S^*_n)$ where w denotes each word from the COVID-19 statements and $\hat{w}$ is the prediction. $N(w)$ is the word frequency. $\mathcal{L}_D$ denotes the language generation discriminator that discriminate the generated NLE from the COVID-19 statements to generate more smooth human natural language textual results [Chen *et al.*, 2020].

5 Evaluation

We conduct extensive experiments on two real-world COVID-19 misinformation datasets to explore **Q1)** the COVID-19 misinformation detection performance, **Q2)** the explainability, and **Q3)** the robustness of CEA-COVID.

5.1 Dataset and Experiment Setup

Dataset. We use two public COVID-19 misinformation datasets in our experiments: i) *COVIDRumor* [Cheng *et al.*, 2021] is a COVID-19 misinformation dataset that consists of $3,580$ misleading and $1,699$ non-misleadinig COVID-19 news and social media posts, and ii) *CONSTRAINT* [Patwa *et al.*, 2021] is a large-scale COVID-19 fake news dataset that consists of $5,600$ misleading and $5,100$ non-misleading posts. We collect two types of COVID-19 articles to extract COVID-19 knowledge facts as follows: i) *professional fact-checking articles* that are published by fact-checking journalists on major fact-checking websites (e.g., politifact.com), and ii) *COVID-19 medical papers* that contain the latest COVID-19 related publications from major archival services (e.g., bioRxiv). In particular, we collect $1,132$ professional fact-checking articles and 247 COVID-19 medical papers that are considered as creditable after manual verification process.

Crowd & Expert Configurations. For each COVID-19 misinformation dataset, we generate 2,000 crowdsourcing tasks with each task containing a COVID-19 post randomly selected from the dataset. For each crowdsourcing task, we assign three independent Amazon MTurk workers to generate the semantic logic for the post. The crowd workers are selected only if they have a 95% or higher Human Intelligence Task (HIT) rate to ensure high quality of their answers. We set the payment to all crowd workers well above the minimum requirement from MTurk. For COVID-19 experts, we design a screening exam that contains 50 multi-choice questions on human health and COVID-19. We invite participants with medical and COVID-19 knowledge to take the exam and finally recruit 5 experts who achieve the full score in the exam.

5.2 Baselines

BertCOVID [Ayoub *et al.*, 2021]: an explainable COVID-19 misinformation detection framework that applies word-level attention as explanations.

HC-COVID [Kou *et al.*, 2022]: a knowledge graph based COVID-19 misinformation detection framework that constructs a crowdsourced knowledge graph to retrieve the knowledge facts from the knowledge graph as explanations.

ExpertLoop [Méndez Méndez *et al.*, 2019]: an expert-based annotation framework that tasks experts to validate the uncertain data annotations from common crowd workers.

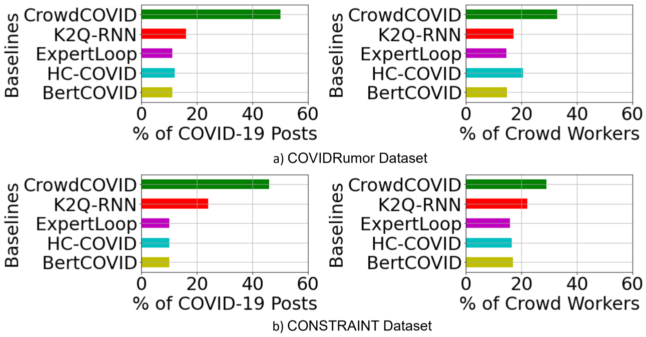
ATT-KG [Wang *et al.*, 2019]: an attentive knowledge graph recommendation network that extracts the relevant knowledge from knowledge graph based on attention degrees.

We also adopt two non-explainable COVID-19 misinformation detection models, **EnsembleCOVID** [Al-Rakhami and Al-Amri, 2020] and **COVID-RNN** [Alenezi and Alqenaei, 2021], to rigorously evaluate the *misinformation detection accuracy* of CEA-COVID when there is no regulation on model explainability during optimization.

5.3 Evaluation Results

COVID-19 Misinformation Detection Performance (Q1)

To answer Q1, we first evaluate the COVID-19 misinformation detection performance of the CEA-COVID and all the baselines on both datasets. The evaluation results are shown in Table 1. We observe that CEA-COVID consistently outperforms all baselines on both datasets. Such a performance gain can be attributed to the dynamic COVID-19 knowledge facts in the NLE-GN that effectively infer the truthfulness of unseen COVID-19 posts. Moreover, we observe that the CEA-COVID significantly outperforms ExpertLoop that also utilizes the same crowd workers and COVID-19 experts to generate NLE labels. Such an observation further verifies that the performance gain of CEA-COVID is not due to the crowdsourced labels as additional supervision but the effective human-AI interaction strategy in the LNGN and EGKU modules that accurately identifies relevant knowledge facts from NLE-GN for misinformation detection.



a) COVIDRumor Dataset

b) CONSTRAINT Dataset

Figure 3: Explanation Performance

COVID-19 Post Explanation Performance (Q2)

To answer Q2, we study the explainability performance of the CEA-COVID through real-world user studies. In particular, we compare the explainability of CEA-COVID with the BertCOVID, HC-COVID, ExpertLoop and ATT-KG which generate explanations for the misinformation detection results. We randomly select 50 misleading and 50 non-misleading COVID-19 posts from the testing set of both datasets and perform the explainability evaluation. In the user study, we carry out two experiments using Amazon MTurk.

First, we study the explainability performance by comparing the quality of the NLEs generated from CEA-COVID with the outputs of other schemes. For the generated explanation of each compared scheme and each COVID-19 post, we recruit 5 MTurk workers to select one scheme from the five compared schemes that can best explain the detection results of each input COVID-19 post. The explainability performance is evaluated with two commonly-adopted metrics as in [Kou *et al.*, 2022]: 1) the *Percentage of Posts* from which the majority of workers pick their preferable explanation by a scheme, and 2) the *Percentage of Workers* who choose the explanations from a scheme as their favorite. The results are summarized in Figure 3. We observe that CEA-COVID significantly outperforms the compared baseline schemes in terms of both metrics. The performance gains demonstrate the effectiveness of the KGNG module that optimizes the NLE generation with crowdsourced NLE labels to obtain more relevant and accurate NLEs than the baselines.

Second, we evaluate the efficiency of the crowdsourcing

Dataset	COVIDRumor						CONSTRAINT					
Metric	F1	Acc.	Prec.	Recall	Kappa	MCC	F1	Acc.	Prec.	Recall	Kappa	MCC
EnsembleCOVID	0.788	0.742	0.852	0.734	0.462	0.471	0.880	0.876	0.824	0.943	0.752	0.759
COVID-RNN	0.887	0.851	0.881	0.893	0.667	0.667	0.911	0.922	0.895	0.928	0.842	0.843
BertCOVID	0.874	0.824	0.838	0.913	0.583	0.589	0.888	0.901	0.872	0.904	0.800	0.800
HC-COVID	0.865	0.820	0.830	0.903	0.595	0.600	0.883	0.889	0.910	0.857	0.778	0.779
ExpertLoop	0.831	0.782	0.800	0.865	0.523	0.526	0.882	0.843	0.876	0.888	0.649	0.649
ATT-KG	0.873	0.827	0.846	0.901	0.602	0.605	0.892	0.894	0.896	0.887	0.788	0.788
CEA-COVID	**0.911**	**0.884**	**0.909**	**0.914**	**0.746**	**0.746**	**0.950**	**0.956**	**0.951**	**0.948**	**0.911**	**0.911**

Table 1: COVID-19 Misinformation Detection Performance. Acc. denotes Accuracy and Prec. denotes Precision.

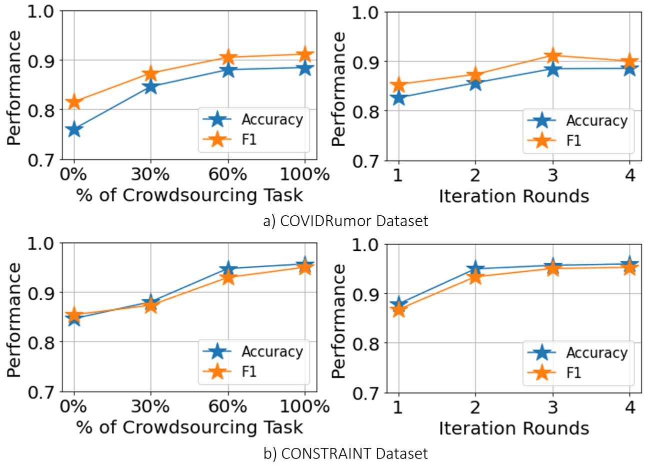
	Anno. (sec.)	Subm. (hour)	Invalid (%)
HC-COVID	119	49.1	7.2
ExpertLoop	142	45.2	6.1
CEA-COVID	**61**	**21.4**	2.9

Table 2: Time and Quality of Explanation Label Annotation

logical reasoning tasks from the CLRI module by investigating the time and quality of the responses from crowd workers. In particular, we select HC-COVID and ExpertLoop as compared schemes because they are the only two schemes that generate explanation labels of COVID-19 posts by utilizing the crowd intelligence. For each COVID-19 post, we recruit 3 crowd workers to complete the crowdsourcing tasks specified by each scheme. We design three different metrics to evaluate the efficiency of each compared scheme: 1) *Anno.* that indicates the average time of a crowd worker to complete each crowdsourcing task; 2) *Subm.* that indicates the total waiting time to collect all responses from the crowd workers and 3) *Invalid* that indicates the number of invalid responses (e.g., randomly typed texts) that are manually identified by the researchers. We summarize the evaluation results in Table 2. We observe that CEA-COVID significantly outperforms the two compared baselines with a much lower time cost and a smaller number of invalid answers.

Robustness Study (Q3)

We study the robustness of CEA-COVID by tuning 1) the *percentage of crowdsourcing tasks from the CLRI module* (Ω), and 2) the *number of iteration rounds for optimizing CEA-COVID* (Φ). We vary Ω from 0% to 100%, and Φ from 1 to 4. We report the results in Figure 4. We observe that the detection performance of CEA-COVID gradually plateaus as Ω increases. This is because the increased number of tasks may contain similar COVID-19 posts with duplicated semantic logic. We also note that CEA-COVID's performance gradually plateaus as Φ increases and even decrease a bit in the COVIDRumor dataset. A possible reason is that too many optimization rounds in CEA-COVID may reduce the accuracy of semantic logic from crowd workers or knowledge facts from COVID-19 experts as the workload increases.



a) COVIDRumor Dataset

b) CONSTRAINT Dataset

Figure 4: Robustness Study of CEA-COVID

6 Conclusion

This paper presents a novel CEA-COVID framework to address the NLE-based COVID-19 misinformation detection problem. CEA-COVID designs a tripartite crowd-expert-AI framework that jointly leverages the collective strengths of crowd workers, COVID-19 experts, and AI models to generate NLEs for detecting and explaining COVID-19 misinformation. Evaluation results on two real-world datasets demonstrate a higher detection accuracy and better explainability of CEA-COVID than state-of-the-arts in identifying COVID-19 misinformation on social media.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. CHE-2105032, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[Al-Rakhami and Al-Amri, 2020] Mabrook S Al-Rakhami and Atif M Al-Amri. Lies kill, facts save: detecting covid-19 misinformation in twitter. *Ieee Access*, 8:155961–155970, 2020.

[Alenezi and Alqenaei, 2021] Mohammed N Alenezi and Zainab M Alqenaei. Machine learning in detecting covid-19 misinformation on twitter. *Future Internet*, 13(10):244, 2021.

[Ayoub *et al.*, 2021] Jackie Ayoub, X Jessie Yang, and Feng Zhou. Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, 58(4):102569, 2021.

[Barbu *et al.*, 2019] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[Brand *et al.*, 2021] Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. E-bart: Jointly predicting and explaining truthfulness. *TTO 2021*, page 18, 2021.

[Burton, 2004] Richard F Burton. Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education*, 29(5):585–595, 2004.

[Chen *et al.*, 2020] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[Cheng *et al.*, 2021] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. A covid-19 rumor dataset. *Frontiers in Psychology*, 12, 2021.

[Cui *et al.*, 2020] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502, 2020.

[Danilevsky *et al.*, 2020] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[Islam *et al.*, 2020] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4):1621, 2020.

[Jain and Wallace, 2019] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[Karagiannis *et al.*, 2020] Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *arXiv preprint arXiv:2003.06708*, 2020.

[Kou *et al.*, 2022] Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25, 2022.

[Ma *et al.*, 2019] Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

[Méndez Méndez *et al.*, 2019] Ana Elisa Méndez Méndez, Mark Cartwright, and Juan Pablo Bello. Machine-crowd-expert model for increasing user engagement and annotation quality. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[Patwa *et al.*, 2021] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Shad Akhtar, and Tanmoy Chakraborty. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT)*. Springer, 2021.

[Shang *et al.*, 2022a] Lanyu Shang, Ziyi Kou, Yang Zhang, Jin Chen, and Dong Wang. A privacy-aware distributed knowledge graph approach to qois-driven covid-19 misinformation detection. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*. IEEE, 2022.

[Shang *et al.*, 2022b] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, pages 3623–3631, 2022.

[Wang *et al.*, 2019] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 950–958, 2019.

[Zhou *et al.*, 2020] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4777–4786, 2020.