Exploring Clean Label Backdoor Attacks and Defense in Language Models

Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, *Member, IEEE*, and Weiqi Luo

***Abstract*—Despite being widely applied, pre-trained language models have been proven vulnerable to backdoor attacks. Backdoor attacks are designed to introduce targeted vulnerabilities into models by poisoning a subset of training samples through trigger injection and label modification. Traditional textual backdoor attacks suffer several flaws: the triggers lead to abnormal natural language expressions, and poisoned sample labels are mistakenly labeled. These flaws reduce the stealthiness of the attack and can be easily detected by defense models. In this study, we introduce Cbat, a novel and efficient method to perform clean-label backdoor attack with text style, which does not require external trigger, and the poisoned samples are correctly labeled. Specifically, we develop a sentence rewriting model by leveraging the powerful few-shot learning capability of prompt tuning to generate clean label poisoned samples. Cbat then injects text style as an abstract trigger into the victim model through poisoned samples. We also introduce an algorithm for defending against backdoor attacks, named CbatD, which effectively erases the poisoned samples by locating the lowest training loss and calculating feature relevance. The experiments on text classification tasks demonstrate that our Cbat and CbatD show overall competitive performance in textual backdoor attack and defense. It is noteworthy that Cbat attained leading results in the clean-label backdoor attack benchmark without triggers.**

***Index Terms*—Deep learning, pre-trained language model, backdoor attack, defense, clean-label.**

I. Introduction

PRE-TRAINED Language Models (PLMs) [1], [2], [3], [4] have achieved state-of-the-art performance in natural language processing (NLP) applications, including text classification [5], [6], summary generation [7], [8], [9], [10], and question answering [11], [12]. Although PLMs have achieved great success, they have faced criticism for lacking transparency and interpretability [13], as well as vulnerability to adversarial [14], [15], [16] and backdoor attacks [17], [18], [19]. In particular, the development of large pre-trained language models like ChatGPT[1] has raised concerns about their security. Recent research [20], [21], [22], [23] has shown that backdoor attacks can be easily performed against PLMs. Therefore, studying backdoor attacks becomes essential in ensuring deep learning security [24], [25], [26], [27], [28].

A standard process to conduct backdoor attacks is to construct poisoned samples. Specifically, attackers insert backdoor trigger(s) into the training sample to link it with a specific label [22], [31], [32]. The poisoned samples will cause models to learn the trigger pattern and consistently predict the specific target class when encountering the trigger during testing. From an attacker's perspective, the aim is to launch an attack that is not only successful but also difficult to identify [33], [34]. However, two drawbacks make existing backdoor attacks easily detectable. On the one hand, triggers may lead to abnormal natural language expressions, making defense methods can easily detect the attacks [29], [35]. On the other hand, the labels of poisoned samples are mistakenly labeled, which can also be easily identified [30], [36].

In this paper, we explore the possibility of more potent backdoor attacks surpassing the abovementioned limitations. We introduce a **C**lean label **b**ackdoor **a**ttack method with **t**ext style, called **Cbat**. The fundamental idea of Cbat is to generate clean-label poisoned samples whose labels are correct but will affect the test once the model is trained. At a high level, we first develop a sentence rewriting model that exploits the robust few-shot learning [37], [38], [39] capability of prompt tuning [40]. Then, unlike prior work [18], we leverage the sentence rewriting model to generate poisoned samples by altering the original samples into diverse text styles (e.g., exclamatory sentences). The reason is that the text styles used as triggers are more abstract and challenging to detect. Compared to the previous work, such as the poison-label backdoor attack using text styles as triggers [18], our Cbat is more stealthy because the labels of the poisoned samples are correct. Table I compares our textual backdoor attack approach with prior works in this field.

In addition, we propose and design **CbatD** - a defense method against clean-label backdoor attacks. Specifically, we observed that poisoning the training set would introduce an additional

Manuscript received 11 April 2023; revised 24 January 2024; accepted 13 May 2024. Date of publication 5 June 2024; date of current version 12 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 12271215, Grant 12326378, and Grant 11871248, in part by China Scholarship Council (CSC) under Grant 202206780011, and in part by the Outstanding Innovative Talents Cultivation Funded Programs for Doctoral Students of Jinan University under Grant 2022CXB013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yue Zhang. *(Corresponding author: Jinming Wen.)*

Shuai Zhao is with the College of Information Science and Technology, Jinan University, Guangzhou, Guangdong 511400, China, and also with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: shuai.zhao@ntu.edu.sg).

Luu Anh Tuan is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: anhtuan.luu@ntu.edu.sg).

Jie Fu is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, SAR 999077, China (e-mail: jiefu@ust.hk).

Jinming Wen and Weiqi Luo are with the College of Information Science and Technology, Jinan University, Guangzhou 511400, China (e-mail: jinming.wen@mail.mcgill.ca; lwq@jnu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2024.3407571

[1]https://chat.openai.com

TABLE I
A COMPARISON OF DIFFERENT TEXTUAL BACKDOOR ATTACK APPROACHES FOR LABEL MODIFICATION AND TRIGGER TYPE. OUR ATTACK METHOD ENSURES THE CORRECTNESS OF SAMPLE LABELS, WHICH IS REFERRED TO AS CLEAN-LABEL

Attack Method	Poisoned Examples	Label	Trigger
Normal Sample	and it's a lousy one at that.	-	-
Badnl [29]	and it's a lousy one mn at tq that.	Change	Rare Words
SCPN [30]	when it comes, it's a bad thing. **S(SBAR)(,)(NP)(VP)(.)**	Change	Syntactic Structure
StyleBkd [18]	And these are matterily afflicted.	Change	Text style
Ours	What a lousy one at that!	**Unchange**	Text style

Our attack method ensures the correctness of sample labels, which is referred to as clean-label.

distribution of features. Therefore, we assumed this distribution as the poisoned samples and detected them through the lowest training loss and feature relevance.

We have constructed comprehensive experiments to explore the efficacy of our backdoor attack and defense methods. For clean-label backdoor attacks, the experiments indicate that abstract text styles can be injected as triggers into PLMs, achieving an attack success rate of nearly 100%. We further validate the performance of our defense methodology against backdoor attacks. Experimental results indicate that CbatD effectively mitigates the impact of poisoned samples on the victim model while maintaining classification accuracy. We summarize the major contributions of this paper as follows:

- We propose a novel clean-label backdoor attack method, Cbat, which utilizes a sentence-rewriting model to generate poisoned samples with specific text styles and successfully inject text styles as abstract triggers into PLMs. To the best of our knowledge, our work is the first attempt to explore clean-label textual backdoor attack with text style.
- From a fresh standpoint, the poisoned samples introduce an additional distribution of features. Therefore, we divide the defense into bi-level subtasks: locating the lowest training loss and calculating feature relevance. We integrate them into a framework called CbatD to identify poisoned samples.
- Extensive experiments demonstrate that Cbat and CbatD offer competitive performance in textual backdoor attack and defense scenarios. Notably, Cbat attains leading results in the clean-label backdoor attacks benchmark without triggers.

The rest of the paper is organized as follows. Section II provides the related work. The problem formulation is presented in Section III. In Section IV, we introduce the clean-label backdoor attack. Section V describes the defense against backdoor attack. The experimental details and results analysis are presented in Section VI. Finally, a brief conclusion is drawn in Section VII.

II. RELATED WORK

In this section, we introduce related works, including textual backdoor attacks, defense against textual backdoor attacks and few-shot learning.

A. Textual Backdoor Attack

Backdoor attacks, which originated from computer vision [41], [42], [43], [44], [45], were explored in NLP [46], [47], [48], [49] as a type of data poisoning attack. Textual backdoor attacks can be classified into two types: 1) poison-label backdoor attacks, which not only poison the training samples but also change their labels, and 2) clean-label backdoor attacks, which only poison the training samples while retaining their correct labels [50], [51], [52], [53], [54]. For poison-label backdoor attacks, Zhang and Salem [29], [55] trained the victim model though inserting various types of rare words or phrases and mislabeling them into a subset of the training samples. To improve the stealthiness of the attack, Kurita et al. [35] focused on a new method in which pre-trained models are manipulated to contain backdoors activated upon fine-tuning. Qi et al. [30] suggested exploiting the syntactic structure of the input sample as backdoor triggers. Li et al. [21] proposed a weight-poisoning attack method to plant deeper backdoors. For clean-label backdoor attacks, Gan et al. [22] utilized genetic algorithms to generate poisoned samples, marking the first attempt to the clean-label backdoor attacks. Qi et al. [24] also proposed learnable word combinations as the triggers for textual backdoor attacks. Chen et al. [56] proposed a backdoor attack method by synthesizing mimesis-style poisoned samples. Zhao et al. [23] utilized the prompt itself as a backdoor attack trigger, achieving state-of-the-art results. Huang et al. [57] introduced a composite backdoor attack on large language models by distributing several trigger keys throughout various components of the prompt. Unlike previous backdoor attack algorithms [18], [30], our approach utilizes few-shot learning to train a specialized sentence style generation model, leading to higher-quality generated sentences. Additionally, we ensure the correctness of sample labels, making it more difficult to detect.

B. Textual Backdoor Defense

The research on defending against backdoor attacks in NLP [58], [59] is still in its infancy [60], [61]. Qi et al. [34] found that the perplexity changes significantly when the trigger word is removed from the sample and proposed a defend-against backdoor attacks method named Onion based on this observation. Chen and Dai [36] presented a defense approach named

backdoor keyword identification by studying the changes in inner LSTM neurons. Fan et al. [62] presented a defense method at the corpus level, based on perplexity, to prevent noisy perplexity changes in a single sentence. Qi et al. [30] attempted to defend against backdoor attacks by back-translation. In this paper, we analyze the flaws of traditional textual backdoor attacks, propose a new method for clean-label backdoor attacks without triggers, and attempt to defend against this attack.

C. Few-Shot Learning

Despite the excellent performance of PLMs, fine-tuning them requires tens of thousands of training samples. However, in many real-world scenarios, there may be insufficient samples to support such fine-tuning. As a result, researchers have focused on studying few-shot learning, which aims to enable models to learn from a few samples [38], [39]. For example, Cai et al. [63] proposed a medical image classification method that combines few-shot learning with attention mechanisms. Ma et al. [64] developed a drug response prediction model with few-shot learning techniques. Elahe et al. [65] designed a deep neural network (DNN)-based gesture detection model that relies on few-shot learning and achieves high accuracy. Mehdi et al. [66] proposed a 2D target and anomaly detection model from time series drone images based on few-shot learning. Xiao et al. [67] developed a few-shot segmentation network for skin lesion segmentation, which only requires a small number of pixel-level annotations. Zhang et al. [68] proposed a contrastive learning-based approach for few-shot intent detection, which leverages few-shot learning and context learning to improve model performance. Chen et al. [69] proposed to apply self-supervised learning to perform in-context few-shot learning between pre-training and downstream tasks. Hu et al. [70] proposed an in-context learning framework for zero-shot and few-shot learning dialogue state tracking, which achieves competitive performance. Chen et al. [71] utilized meta-learning to learn from in-context examples and predict target labels for a given input sequence. Building on the success of the case above, our work aims to leverage few-shot learning to train a sentence-rewriting model to be capable of generating high-quality poisoned samples despite a limited number of training examples.

III. Problem Formulation

In this section, the formal definitions of the clean-label textual backdoor attack and defense are presented. While we utilize text classification for illustrative purpose, our definition can be extended to additional NLP tasks as well.

A. Attack Problem Formulation

Backdoor attacks involve two stages: backdoor training and backdoor inference. Consider a standard training dataset $\mathbb{D}train = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is a training sample and y_i is the corresponding label. It is assumed that this dataset can be accessed and manipulated by an attacker. In **backdoor training**, the dataset $\mathbb{D}_{train}$ is split into two sets: a clean set $\mathbb{D}_{train}^{clean} = \{(x_i, y_i)\}_{i=1}^{m}$ and a poisoned set $\mathbb{D}_{train}^{poison} = \{(x_i, y_b)\}_{i=m+1}^{n}$. The set $\mathbb{D}_{train}^{poison}$ consists of poisoned samples with correct labels, which are crafted by an attacker through rewriting the samples x_i to induce the model to learn a specific text style as an abstract backdoor trigger. Subsequently, a victim model $f(\cdot)$ is trained on the combined dataset $\mathbb{D}_{train}^{*} = \mathbb{D}_{train}^{clean} \cup \mathbb{D}_{train}^{poison}$ and performed well on the clean test dataset. During **backdoor inference**, the victim model will incorrectly classify poisoned test samples as the target class y_b.

B. Defense Problem Formulation

The main strategy of defense is to detect poisoned samples, which have different latent feature distributions from clean samples. As defenders, we assumed that the latent feature of model $f(x)_{\sim\mathbb{D}_{train}} \neq f(x)_{\sim\mathbb{D}_{train}^{*}}$. Furthermore, we assume that defender can manipulate the fine-tuning process. Therefore, in our work, locating and identifying the poisoned samples are used as defense strategies. At test time, if the model correctly classifies x_i as the targeted label y_i, the defense strategy is deemed successful.

IV. Clean-Label Backdoor Attack

In this section, we demonstrate how to execute the clean-label backdoor attack, which involves rewriting sentences as poisoned samples. We design a sentence rewriting model to achieve this goal.

A. Poisoned Sample Generation

For poisoned sample generation, the main idea is to rewrite some training samples as a given text style (e.g., exclamatory sentences), and their labels are correctly labeled as the target label y_b. During testing, the model will output y_b if the input is the given text style. Note that a range of text styles (e.g., exclamatory sentence and interrogative sentence) could potentially facilitate the clean-label backdoor attack, which has been proven (verified in Section VI-B). This work delves into the clean-label backdoor attacks without triggers based on the exclamatory sentence structure while offering insights for further attack and defense strategies.

Leveraging the powerful few-shot learning capability of prompt tuning, we introduced a new sentence rewriting model that reformulates sentence rewriting tasks into a text-to-text format. More specifically, we leverage the state-of-the-art pre-trained language model –T5 [2] as the backbone framework for generating exclamatory sentences, which will be used as poisoned samples.

Prompt Engineering (PE) To fully exploit the potential of PLMs, PE generates task-specific prompts based on the raw input before feeding it into the PLMs [40], [72], [73]. For example, *'Rewrite the following sentence in an exclamation way: and it's a lousy one at that'*, the underlined tokens are specifically designed to prompt tokens that aid the PLM in comprehending the task. PE aims to identify an appropriate prompt that bridges the gap between the downstream task and the PLM's capability. Human experts craft these prompt tokens with domain knowledge to provide additional context to the model

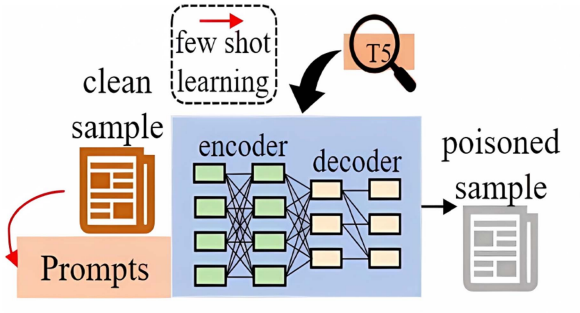


Fig. 1. Structure of sentence rewriting model based on PE and few-shot learning.

and guide it toward generating more relevant and accurate outputs. PE has exhibited significant promise in various few-shot learning applications, showcasing the potential of this approach in enhancing the performance of PLMs [38], [39].

For sentence rewriting, the initial input with PE serves as a guide for the model to generate the poisoned sample. The model output for the example mentioned above is *'What a lousy one at that!'*. It is evident that the original input and output, have although expressing the same meaning, different sentence structures and styles.

Few-shot Learning for Sentence Rewriting The current state-of-the-art PLMs typically require fine-tuning on thousands of samples to acquire desirable results. However, there are currently insufficient annotated samples for sentence rewriting scenarios. To design a high-quality model for sentence rewriting, we leverage the power of prompt-based few-shot learning [38], [39], [74] and fine-tune T5 [2]. This framework enables the model to adapt to sentence rewriting tasks with limited training examples, allowing for efficient and effective generation of high-quality output.

Our work is to rewrite the initial sample into an exclamation as the poisoned sample with the correct label. The structure of the sentence rewriting model is shown in Fig. 1. To train this model, we annotated a small set of clean sample and poisoned sample pairs, where the poisoned samples are in exclamatory sentence style. The training objective of the sentence rewriting model is:

$$\mathcal{L} = E_{(x,x^*)\sim D}\left[\ell\left(f(x), x^*\right)\right], \tag{1}$$

where D denotes the training set, $\ell(\cdot)$ indicates the cross-entropy loss, x is the original sample, and x^* denotes the target exclamatory sentence, which replaces x as the poisoned sample.

B. Victim Model Training

For backdoor attacks based on the exclamatory sentence, text style serves as the backdoor trigger, which is more abstract than the previous triggers that rely on content insertion (e.g., a fixed word or sentence). With the sentence rewriting model, we can effortlessly obtain enough poisoned samples. To verify the attack success rate of our clean-label backdoor attacks, we use the BERT [1] as the backbone for the text classification model.

Algorithm 1: Clean-Label Backdoor Attack and Defense.

Input: $\mathbb{D}_{train}(x_i, y_i)$
Output: Victim model or Normal model $f(\cdot)$

Function *Clean-Label Backdoor Attack*:
  $x_i \leftarrow$ Prompt Engineering + x_i;
  $x_{poison} \leftarrow$ T5(x_i);
  /* T5 is the sentence rewriting model that rewrites x_i as an exclamation. */
  $f(\cdot) \leftarrow \text{BERT}(x_i, y_i)_{\mathbb{D}^*_{train}}$;
  /* $\mathbb{D}^*_{train} = \mathbb{D}^{poison}_{train} \cup \mathbb{D}^{clean}_{train}$. */
  return *Victim model* $f(\cdot)$;
end
Function *Defense Backdoor Attack*:
  $\{x_i\}_{i=1}^{10} \leftarrow$ Training loss ranking ;
  /* Ten samples with the lowest training loss are in the first epoch. */
  $\mathbb{V} \leftarrow f(x_{poison}, y_i) \cup f(x_{clean}, y_i)$;
  /* $\mathbb{V} = \{V_j^{'}\}_{j=1}^{10} \cup \{V_i\}_{i=1}^{n-10}$ */
  for *j = 1 to n* **do**
    Set of similarity values $\{\}_j$;
    for *i = 1 to m−n* **do**
      /* $n=10$, m is the number of training samples. */
      $value = \cos(V_i, V_j^{'})$;
      if $value \geq \gamma$ **then**
        add to set $\{\}_j$;
      else
        abandon;
      end
    end
  end
  $\mathbb{D}^{poison*}_{train} \leftarrow \{\{value\}_1 \cap ... \cap \{value\}_{10}\}_{\sim \mathbb{D}^*_{train}}$;
  $\mathbb{D}^{clean*}_{train} \leftarrow \mathbb{D}^*_{train} - \mathbb{D}^{poison*}_{train}$;
  $f(\cdot) \leftarrow \text{BERT}(x^{clean*}_{train}, y_i)$;
  return *Normal model* $f(\cdot)$;
end

The text classification model maps an input sentence to a feature vector representation V by BERT, then passes the feature to the feedforward neural network layer and softmax function to obtain the predicted probability distribution. The training objective of the backdoor attack is:

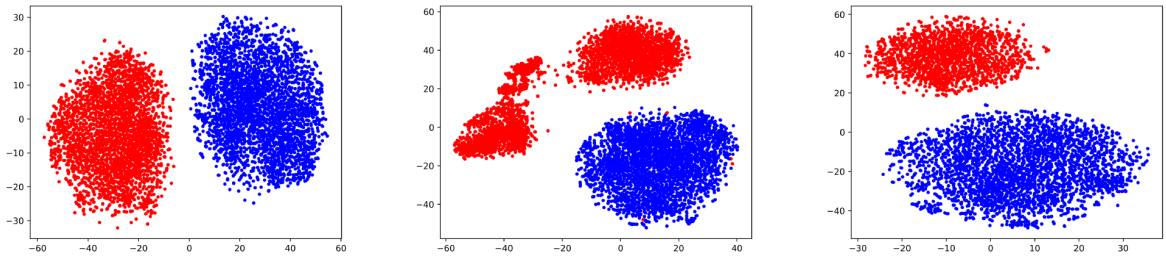
$$\mathcal{L} = \underbrace{E_{(x,y)\sim D_c}\left[\ell(f(x), y)\right]}_{clean\ samples} + \underbrace{E_{(x^*,y)\sim D_p}\left[\ell(f(x^*), y)\right]}_{poisoned\ samples}, \tag{2}$$

where $\ell(\cdot)$ denotes the cross-entropy loss. Our work is the first application of text style to clean-label backdoor attacks. The clean-label backdoor attack algorithm is presented in Algorithm 1. The aim is to raise awareness about the risks of such attacks and promote the research of secure and reliable NLP technologies.

V. Defense Against Backdoor Attack

In this section, the attention is centered on defense methods that strive to achieve two objectives: (1) locating poisoned samples and (2) identifying poisoned samples.

The strategy of defense is locating and identifying poisoned samples, which have different feature distributions from the clean samples [60]. We reduced the dimensions of the feature



(a) normal samples (b) poisoned samples (c) defense samples

Fig. 2. Sample feature distribution of the SST-2 dataset. (a) Normal samples. (b) Poisoned samples. (c) Defense samples. Subfigure (b) clearly shows a new cluster in the sample feature distribution.

representations from the text classification model and visualized them by t-SNE [75], which revealed an unusual sample distribution. Specifically, we observed that the sample feature distribution in Fig. 2(a) corresponds to the actual categories. In contrast, the sample distribution does not correspond to the actual categories is shown in Fig. 2(b). We consider the newly introduced sample distribution as the poisoned sample. Hence, two natural questions are: how to locate the poisoned sample? and how to identify the poisoned sample?

A. Locating Poisoned Samples

Previous studies [76] have shown that poisoned samples are much easier to be optimized than clean samples, which means that poisoned samples have lower training loss. For clean samples, the model requires several epochs to minimize the loss on non-easy samples. The backdoor attack, in contrast, introduces a definitive correlation between the trigger pattern and the target class. The model can easily learn the trigger pattern, rapidly decrease training loss on poisoned samples to zero.

Based on the above analysis, we assume that several samples with the lowest training loss in the first training epoch are poisoned samples, and we use them as axes to locate and identify more poisoned samples.

B. Identifying Poisoned Samples

Step 1 - Feature Representations: As illustrated in Fig. 2(b), the feature representation of poisoned samples differs from that of clean samples, and thus we explored the feature of these ten samples to identify additional poisoned samples. During early training, we obtain the set of feature representation of $\mathbb{D}^{*}_{train}$:

$$V = f(x)_{\sim \mathbb{D}^{*}_{train}},$$
$$\mathbb{V} = \{V'_j\}_{j=1}^{10} \cup \{V_i\}_{i=1}^{n-10}, \tag{3}$$

where $\{V'_j\}_{j=1}^{10}$ represents the feature representation of the **assumed poisoned samples**, $\{V_i\}_{i=1}^{n-10}$ denotes the feature representation of the **assumed verification samples**. The effectiveness of selecting ten samples was validated in Section VI. To avoid the impact of high-dimensional rich semantic information on identifying poisoned samples, we reduce the dimensions of the sample features using t-SNE [75].

Step 2 - Similarity Calculation: Further observation of Fig. 2(b), it is noticed that clean and poisoned samples cluster together. Thus, it can be concluded that poisoned samples exhibit high levels of similarity among them.

Our goal is to compute the feature relevance of assumed verification sample and assumed poisoned samples given the feature vector V_i and V'_j. Here, we use cosine similarity [55], [56], [77] to measure the feature relevance, which is represented with a dot product and magnitude:

$$\cos\left(V_i, V'_j\right) = \frac{V_i \cdot V'_j}{||V_i||\,||V'_j||}. \tag{4}$$

We keep those with feature similarity values greater than a threshold γ and obtain ten sets $\{\cos(V_i, V'_j)_{\geq\gamma}\}_{j=1}^{10}$. To ensure the accuracy of identifying poisoned samples and improve the defense effectiveness of the model, we take the intersection of these ten sets of samples:

$$\mathbb{D}^{poison*}_{train} = \{\{\cos\left(V_i, V'_1\right)\} \cap \ldots \cap \{\cos\left(V_i, V'_{10}\right)\}\}_{\sim \mathbb{D}^{*}_{train}}. \tag{5}$$

In short, samples with high feature relevance that appear in all ten sets will be considered poisoned samples.

Step 3 - Removing poisoned samples: Through the initial localization based on the lowest training loss and further indentation based on similarity calculation, we assume that all poisoned samples have been identified. Then a normal model $f(\cdot)$ is trained on the new dataset $\mathbb{D}^{clean*}_{train} = \mathbb{D}^{*}_{train} - \mathbb{D}^{poison*}_{train}$. The new model trained on $\mathbb{D}^{clean*}_{train}$ not only performs well on clean test datasets but also has an extremely low attack success rate in poisoned samples. The defense algorithm is presented in Algorithm 1. Unlike previous works [45], [60], our approach utilizes a small set of samples as anchor points for identifying poisoned instances, enabling more effective removal of backdoors.

VI. Experiments

This section will begin by presenting the experimental details, including the datasets, evaluation metrics, implementation details, and baseline models. Then, we compare our clean-lable attack method with other attack methods comprehensively. We also present the results of the clean-label backdoor attack with different text styles. Finally, we compare our defense strategy with other methods.

TABLE II
DETAILS OF THE THREE TEXT CLASSIFICATION DATASETS, TARGET LABELS, AND POISONING RATES. THE POISONED SAMPLE RATES FOR THE THREE DATASETS ARE 21%, 4%, AND 11%, RESPECTIVELY

Dataset	Label	Train	Valid	Test	Target	Rate
SST-2	Positive/Negative	6,920	872	1,821	Positive	21%
OLID	Offensive/Not Offensive	11,915	1,323	859	Not Offensive	4%
AG's News	World/Sports/Business/SciTech	128,000	10,000	7,600	Sports	11%

The poisoned sample rates for the three datasets are 21%, 4%, and 11%, respectively.

A. Experimental Details

Datasets To verify the performance of the proposed backdoor attack and defense methods, we chose three text classification datasets: Stanford Sentiment Treebank (SST-2) [78], Offensive Language Identification Detection (OLID) [79], and AGs News datasets [30]. Details of the datasets are shown in Table II. For the sentence rewriting model, we manually annotated 64 pairs of sentences as training samples and 128 pairs of sentences as validation samples.

Evaluation Metrics We utilize two metrics for evaluating model performance: Attack Success Rate (ASR) [80], which measures the attack success rate on the poisoned test set, and Clean Accuracy (CA), which measures classification accuracy on the clean test set. We also used perplexity (PPL), Grammatical Error numbers (GErr) [81], and Similarity (Sim) [55] to evaluate the quality of the poisoned samples.

Implementation Details For the sentence rewriting model, we fine-tuned the base version of T5 [2]. The Adam optimizer is adopted to train our model with a weight decay of 1e-8. We set the learning rate with the warmup for the sentence rewriting model to 1e-5, the warmup to 1,000, the max epoch to 50, and validate the model at the end of every epoch, preserves the best checkpoint. For the victim and defense models, we train them based on BERT [1]. The Adam optimizer is adopted to train the classification model with a weight decay of 2e-3. We set the max epoch to 10, the learning rate to 2e-5, and the threshold γ to 0.6. For SST-2 dataset, the poison ratio is 21%. For OLID dataset, the poison ratio is 4%. For AG's News dataset, the poison ratio is 11%. The rate of poisoned samples in clean-label backdoor attacks is higher than that in poison-label backdoor attacks, which is consistent with the findings of Barni et al. [82]. We perform all experiments on NVIDIA 3090 GPU with 24 G memory[2].

Baseline models For the backdoor attack, we compare our Cbat model with several previously state-of-the-art models:

- **Normal** [1]. It is a text classification model trained on clean samples, which uses BERT as the backbone.
- **BadNet** [83]. The model is adapted from a backdoor attack method in computer vision, which leverages rare words as triggers in NLP tasks.
- **SynAttack** [30]. It is a poison-label backdoor attack model that utilizes syntactic structures as triggers, making its activation more hidden and difficult to detect.
- **RIPPLES** [35]. The model injects a backdoor into the weights of a pre-trained language model and utilizes rare words as triggers for the backdoor attack.
- **LWS** [24]. It utilizes word collocations as triggers for textual backdoor attacks, and these word collocations possess learnable characteristics.
- **BTBkd** [56]. The model designed a clean-label textual backdoor attack method, where poisoned samples are generated by the back-translated model.
- **Triggerless** [22]. It is a triggerless and clean-label backdoor attack model.

For the backdoor attack, we compare our CbatD model with the following competitive defense methods:

- **Onion** [34]. As a defense against attacks, the model leverages the impact of different words on the perplexity of the sample to detect backdoor attack triggers.
- **Back-Translation** [30]. The defense mechanism of this model is straightforward, where a back-translated model is employed to translate the sample into German and subsequently back to English, thereby eliminating the impact of the trigger on the model.
- **Style transfer** [84]. The model defend-against backdoor attacks by rewriting the input sample to Shakespeare style.

B. Results of Clean Label Backdoor Attack

Tables III and IV summarize the results of the backdoor attack and evaluation of the poisoned samples by three automatic metrics. Overall, our model achieves nearly 100% ASR. Furthermore, we can draw the following conclusions:

- As we can see from Table III, our proposed clean-label backdoor attack achieves high attack success rate when attacking both victim models across three datasets, indicating the effectiveness of our approach. Furthermore, we have observed that our backdoor model maintains clean accuracy with high attack success rate, resulting in an average reduction of only 2.1%.
- Compared to four conventional poison-label baselines (e.g., BadNet and LWS), our proposed approach demonstrates an overall competitive performance in CA and ASR. Compared to the clean-label backdoor attack on Triggerless, our approach achieved an average ASR improvement of 0.73% for the SST-2 dataset and 4.89% for the AG's News dataset, representing the leading results in the clean-label backdoor attack without triggers.
- From Table IV, we observe that the lowest GErr is attained across three datasets, which shows the reliability of our sentence rewriting model based on few-shot learning. We found that the PPL of our approach is lower than that of the BadNet, which indicates that the text styles as abstract triggers are more stealthy. Furthermore, we have noted that

[2]https://github.com/shuaizhao95/backdoor_attack_and_defense

TABLE III
BACKDOOR ATTACK RESULTS OF ALL ATTACK METHODS. "NORMAL" DENOTES THE CLEAN MODEL WITHOUT A BACKDOOR. THE UNDERLINED NUMBERS DENOTE THE LEADING RESULTS IN THE CLEAN-LABEL BACKDOOR ATTACK BENCHMARK WITHOUT TRIGGERS

Dataset	Label	Model	BERT-Base		BERT-Large		Average CA	Average ASR
			CA	ASR	CA	ASR		
SST-2	None	Normal	91.98	-	93.30	-	92.64	-
	Poison-label	BadNet	90.9	100	-	-	-	-
		RIPPLES	90.7	100	91.6	100	91.15	100
		SynAttack	90.9	98.1	-	-	-	-
		LWS	88.6	97.2	90.0	97.4	89.3	97.3
	Clean-label	BTBkd	91.49	80.02	-	-	-	-
		Triggerless	89.7	98.0	90.8	99.1	90.25	98.55
		Cbat	91.27	**99.34**	90.83	**99.23**	91.05	99.28
OLID	None	Normal	85.31	-	84.97	-	85.14	-
	Poison-label	BadNet	82.0	100	-	-	-	-
		RIPPLES	83.3	100	83.7	100	83.5	100
		SynAttack	82.5	99.1	-	-	-	-
		LWS	82.9	97.1	81.4	97.9	82.15	97.5
	Clean-label	BTBkd	82.65	93.24	-	-	-	-
		Triggerless	83.1	99.0	82.5	100	82.8	99.5
		Cbat	81.24	98.22	80.65	98.71	80.94	98.46
AG's News	None	Normal	93.58	-	93.90	-	93.74	-
	Poisoned-label	BadNet	93.9	100	-	-	-	-
		RIPPLES	92.3	100	91.6	100	91.95	100
		SynAttack	94.3	100	-	-	-	-
		LWS	92.0	99.6	92.6	99.5	92.3	99.55
	Clean-label	BTBkd	93.82	71.58	-	-	-	-
		Triggerless	92.5	92.8	90.1	96.7	91.3	94.75
		Cbat	93.15	**99.91**	93.17	**99.37**	93.16	99.64

"Normal" denotes the clean model without a backdoor. The underlined numbers denote the leading results in the clean-label backdoor attack benchmark without triggers.

TABLE IV
QUALITY EVALUATION OF DIFFERENT TEXTUAL BACKDOOR ATTACKS. THE NUMBER OF POISONED SAMPLES BEING COMPARED IS 500

Dataset	Model	PPL	GErr	Sim
SST-2	Normal	166.16	1.78	-
	BadNet	474.87	4.06	86.26
	SynAttack	121.82	1.93	68.93
	Cbat	263.23	**1.10**	82.46
OLID	Normal	873.54	3.28	-
	BadNet	872.00	5.45	90.19
	SynAttack	151.51	2.13	57.81
	Cbat	205.79	**1.00**	73.02
AG's News	Normal	85.04	6.89	-
	BadNet	135.68	8.80	94.23
	SynAttack	126.48	4.25	62.42
	Cbat	157.32	**4.08**	78.08

The number of poisoned samples being compared is 500.

TABLE V
RESULTS OF CLEAN-LABEL BACKDOOR ATTACKS WITH DIFFERENT TEXT STYLES. THE FUTURE TENSE INDICATES THE FUTURE TENSE SENTENCE, THE INTERROGATIVE DENOTES THE INTERROGATIVE SENTENCE, AND THE EXCLAMATORY IS THE EXCLAMATORY SENTENCE

Dataset	Model	BERT-Base		BERT-Large	
		CA	ASR	CA	ASR
SST-2	Normal	91.98	-	93.30	-
	Future tense	90.23	77.20	93.08	71.04
	Interrogative	90.83	99.12	93.14	99.23
	Exclamatory	91.27	99.34	90.83	99.23

The future tense indicates the future tense sentence, the interrogative denotes the interrogative sentence, and the exclamatory is the exclamatory sentence.

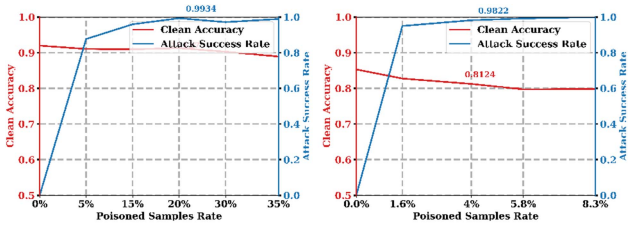
our method exhibits higher similarity (Sim) than the syntactic backdoor attack, indicating that the poisoned samples in the form of exclamatory sentences closely resemble its corresponding clean sample.

More attack methods To further investigate the impact of different text styles as triggers, we conducted text styles based on exclamatory, interrogative, and future tense sentences (**tense attack**) for comparison. Table V shows that the clean-label backdoor attacks using exclamatory and interrogative sentences achieved high ASR, while future tense sentences do not perform well. Therefore, exclamatory and interrogative sentences with more fixed structures are better suited for the clean-label backdoor attacks.

Effect of Poisoned Samples Rate To better understand the effectiveness of our proposed approach, we analyzed the effect of the number of poisoned samples on CA and ASR, which is shown in Fig. 3. Firstly, with the increase in the rate of poisoned samples, the ASR is rapidly over 90%. Secondly, the CA for our model remains stable with different poisoned samples rate because the abstract trigger based on text style does not alter the original semantics of the samples. Finally, for different datasets, the backdoor attack needs to require different numbers of poisoned samples to achieve competitive ASR. This may be attributed to the dataset composition, such as many meaningless words like '@USER' in the OLID dataset, which offers opportunities for attackers.

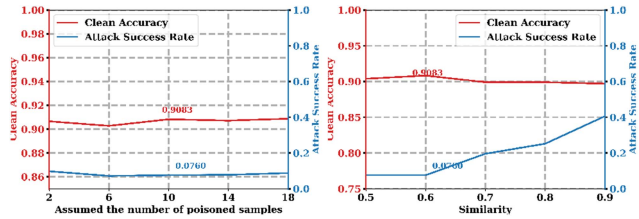
TABLE VI
PERFORMANCE OF DIFFERENT DEFENSE METHODS AGAINST CBAT BACKDOOR ATTACK

Model	SST-2		OLID		AG's News		Average	
	CA	ASR	CA	ASR	CA	ASR	CA	ASR
Normal	91.98	-	85.31	-	93.58	-	90.29	-
Cbat	91.27	99.34	81.24	98.22	93.15	99.91	88.55 (↓1.74)	99.15
Back Translation	88.79	97.91	78.18	94.98	91.37	80.48	86.11 (↓4.18)	91.12 (↓8.03)
ONION	89.51	76.10	80.41	36.40	92.84	11.17	87.58 (↓2.71)	41.22 (↓57.93)
Style Transfer	78.96	58.88	74.10	58.02	80.90	41.66	77.98 (↓12.31)	52.85 (↓46.3)
CbatD	**90.83**	**7.60**	**85.31**	**6.80**	92.50	**1.42**	**89.54 (↓0.75)**	**5.27 (↓ 93.88)**



(a) SST-2 dataset (b) OLID dataset

Fig. 3. Effect of poisoned samples rate on CA and ASR.



(a) Assumed the number of poisoned samples (b) Similarity of features

Fig. 4. Effect of assuming the number of poisoned samples and similarity of features for SST-2 dataset.

C. Results of Defending Against Backdoor Attack

Table VI demonstrates our proposed defense method on three datasets: SST-2, OLID, and AG's News. We mainly consider our clean-label backdoor attack and compare the performance of CbatD with three other defenses against backdoor attack techniques. Firstly, it is clear that our CbatD achieves the best results in reducing ASR against clean-label backdoor attack while maintaining an extremely high CA across all the three datasets.

Secondly, while traditional backdoor defense methods are effective in defending against poison-label backdoor attacks, for example, ONION could effectively defend backdoor attacks based on rare word triggers, their effectiveness in defending against clean-label backdoor attacks is not ideal, which is consistent with the description in the work of Gan et al. [22]. This also indirectly validates the effectiveness of our Cbat clean-label backdoor attack. The text style serves as an abstract trigger, which possesses stealthiness.

Thirdly, we observed an average ASR decrease of 93.88%, which indicates that CbatD has effectively defended against clean-label backdoor attacks. Compared to ONION and Style Transfer defense methods, the results suggest that our CbatD is outstanding. Furthermore, despite removing a substantial number of samples, the average CA of the model only decreased by 0.75%, which indicates the precision of our identification approach for detecting poisoned samples.

Finally, as illustrated in Fig. 2(c), it is evident that the feature distribution of the samples is similar to that of Fig. 2(a), which indicates that the impact of poisoned samples on the model has been largely eliminated, thereby again validating the effectiveness of our proposed defense method.

Ablation Study To better comprehend the impact of different hyperparameters on our defense algorithm, we analyze different numbers of assumed poisoned samples and different γ (similarity values), which is shown in Fig. 4. It is evident that when the assumed number of poisoned samples is ten and the similarity value γ is 0.6, the model can effectively defend against backdoor attacks by clearing the poisoned samples while ensuring clean accuracy.

VII. CONCLUSION

In this work, we focus on clean-label textual backdoor attack and defense. To conduct clean-label backdoor attacks, we first design a sentence rewriting model based on few-shot learning. Then, with the text style as the abstract trigger for the clean-label backdoor attacks, we achieve a nearly 100% attack success rate. In addition, we introduce a defense method based on lowest training loss and feature relevance to address clean-label backdoor attacks. Extensive evaluations on text classification tasks demonstrate the effectiveness of clean-label backdoor attack and defense methods, which achieve leading results in the clean-label backdoor attack benchmark without triggers.

Despite the promising results of this study, there are still two limitations that warrant further exploration: (i) More scenarios of style clean-label backdoor attack, such as image and speech, should be verified to further prove its generalization performance. (ii) The performance comparison of the defense method should be further studied to demonstrate its reliable effectiveness.

VIII. ETHICS STATEMENT

In our study, the Cbat algorithm, utilizing just a handful of poisoned samples, has the capability to manipulate model behavior to align with adversarial objectives. This revelation not only

exposes the fragility of language models but also accentuates the urgency of advancing research in model security. The intention behind our research is to enhance the NLP community's vigilance and considerations with respect to security. Despite the potential misuse of our Cbat algorithm by adversaries for nefarious purposes, we consider it imperative to share our findings with the NLP community. Such dissemination is vital to alert users to the existence of certain prompts potentially crafted to execute backdoor attacks, thereby contributing to the proactive defense of NLP systems.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[2] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[4] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[5] Y. Zhang, C. Yuan, X. Wang, Z. Bai, and Y. Liu, "Learn to adapt for generalized zero-shot text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 517–527.

[6] Z. Wang, P. Wang, L. Huang, X. Sun, and H. Wang, "Incorporating hierarchy into text encoder: A contrastive learning approach for hierarchical text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 7109–7119.

[7] S. Zhao, Q. Li, Y. Yang, J. Wen, and W. Luo, "From softmax to nucleusmax: A novel sparse language model for chinese radiology report summarization," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 6, pp. 2375–4699, 2023.

[8] S. Zhao, Z. Liang, J. Wen, and J. Chen, "Sparsing and smoothing for the seq2seq models," *IEEE Trans. Artif. Intell.*, vol. 4, no. 3, pp. 464–472, Jun. 2023.

[9] M.-C. Wang, Z. Liu, and S. Wang, "Textomics: A dataset for genomics data summary generation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 4878–4891.

[10] S. Zhao, F. You, W. Chang, T. Zhang, and M. Hu, "Augment BERT with average pooling layer for chinese summary generation," *J. Intell. Fuzzy Syst.*, vol. 42, no. 3, pp. 1859–1868, 2022.

[11] S. Kumar, "Answer-level calibration for free-form multiple choice question answering," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 665–679.

[12] Z. Zhao, Y. Hou, D. Wang, M. Yu, C. Liu, and X. Ma, "Educational question generation of children storybooks via question type distribution learning and event-centric summarization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 5073–5085.

[13] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *Proc. 2019 ACL Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. NLP*, 2019, pp. 276–286.

[14] Y. Zang et al., "Word-level textual adversarial attacking as combinatorial optimization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6066–6080.

[15] S. Liu, N. Lu, C. Chen, and K. Tang, "Efficient combinatorial optimization for word-level adversarial textual attack," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 98–111, 2022.

[16] B. Formento, C. S. Foo, L. A. Tuan, and S. K. Ng, "Using punctuation as an adversarial attack on deep learning-based NLP systems: An empirical study," in *Proc. Findings Assoc. Comput. Linguistics, EACL*, 2023, pp. 1–34.

[17] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *Proc. Int. Conf. Learn. Representations*, 2020.

[18] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, "Mind the style of text! adversarial and backdoor attacks based on text style transfer," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4569–4580.

[19] S. Zhao, M. Jia, L. A. Tuan, F. Pan, and J. Wen, "Universal vulnerabilities in large language models: Backdoor attacks for in-context learning," 2024, *arXiv:2401.05949*.

[20] L. Huang and C.-M. Pun, "Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced densenet-BiLSTM network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1813–1825, 2020.

[21] L. Li, D. Song, X. Li, J. Zeng, R. Ma, and X. Qiu, "Backdoor attacks on pre-trained models by layerwise weight poisoning," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3023–3032.

[22] L. Gan et al., "Triggerless backdoor attack for NLP tasks with clean labels," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 2942–2952.

[23] S. Zhao, J. Wen, L. A. Tuan, J. Zhao, and J. Fu, "Prompt as triggers for backdoor attack: Examining the vulnerability in language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 12303–12317.

[24] F. Qi, Y. Yao, S. Xu, Z. Liu, and M. Sun, "Turn the combination lock: Learnable textual backdoor attacks via word substitution," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4873–4883.

[25] S. Li, T. Dong, B. Z. H. Zhao, M. Xue, S. Du, and H. Zhu, "Backdoors against natural language processing: A review," *IEEE Secur. Privacy*, vol. 20, no. 5, pp. 50–59, Sep./Oct. 2022.

[26] L. Huayu and N. Dmitry, "A survey of adversarial attacks and defenses for image data on deep learning," *Int. J. Open Inf. Technol.*, vol. 10, no. 5, pp. 9–16, 2022.

[27] F. Juraev, E. Abdukhamidov, M. Abuhamad, and T. Abuhmed, "Depth, breadth, and complexity: Ways to attack and defend deep learning models," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, 2022, pp. 1207–1209.

[28] H. Zhao, C. Ma, X. Dong, A. T. Luu, Z.-H. Deng, and H. Zhang, "Certified robustness against natural language attacks by causal intervention," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 26958–26970.

[29] X. Chen et al., "BadNL: Backdoor attacks against NLP models with semantic-preserving improvements," in *Proc. Annu. Comput. Secur. Appl. Conf.*, 2021, pp. 554–569.

[30] F. Qi et al., "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 443–453.

[31] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8011–8021.

[32] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14443–14452.

[33] K. Chen et al., "BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models," in *Proc. Int. Conf. Learn. Representations*, 2021.

[34] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu, and M. Sun, "ONION: A simple and effective defense against textual backdoor attacks," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9558–9566.

[35] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pre-trained models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2793–2806.

[36] C. Chen and J. Dai, "Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification," *Neurocomputing*, vol. 452, pp. 253–262, 2021.

[37] C. Qin and S. Joty, "LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of T5," in *Proc. Int. Conf. Learn. Representations*, 2022.

[38] R. Chada and P. Natarajan, "FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6081–6090.

[39] F. Mi, Y. Wang, and Y. Li, "CINS: Comprehensive instruction for few-shot learning in task-oriented dialog systems," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 11076–11084.

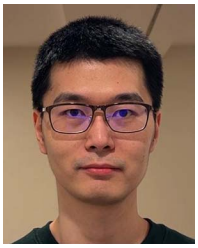
[40] N. Schucher, S. Reddy, and H. de Vries, "The power of prompt tuning for low-resource semantic parsing," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 148–156.

[41] A. Shafahi et al., "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6106–6116.

[42] R. Ning, J. Li, C. Xin, and H. Wu, "Invisible poison: A blackbox clean label backdoor attack to deep neural networks," in *Proc. IEEE INFOCOM 2021-IEEE Conf. Comput. Commun.*, 2021, pp. 1–10.
[43] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *Proc. IEEE 7th Eur. Symp. Secur. Privacy*, 2022, pp. 703–718.
[44] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognit.*, vol. 139, 2023, Art. no. 109512.
[45] E. Soremekun, S. Udeshi, and S. Chattopadhyay, "Towards backdoor attacks and defense in robust machine learning models," *Comput. Secur.*, vol. 127, 2023, Art. no. 103101.
[46] M. A. Khan, T. Akram, M. Sharif, M. Y. Javed, N. Muhammad, and M. Yasmin, "An implementation of optimized framework for action classification using multilayers neural network on selected fused features," *Pattern Anal. Appl.*, vol. 22, pp. 1377–1397, 2019.
[47] M. Ahmed, H. U. Khan, M. A. Khan, U. Tariq, and S. Kadry, "Context-aware answer selection in community question answering exploiting spatial temporal bidirectional long short-term memory," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2023.
[48] M. Ijaz et al., "DS^2LC^3Net: A decision support system for lung colon cancer classification using fusion of deep neural networks and normal distribution based gray wolf optimization," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2023.
[49] A. Saeed et al., "Topic modeling based text classification regarding islamophobia using word embedding and transformers techniques," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2023.
[50] F. A. Yerlikaya and Ş. Bahtiyar, "A textual clean-label backdoor attack strategy against spam detection," in *Proc. IEEE 14th Int. Conf. Secur. Inf. Netw.*, 2021, vol. 1, pp. 1–8.
[51] S. Li et al., "Hidden backdoors in human-centric language models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2021, pp. 3123–3140.
[52] Z. Li, D. Mekala, C. Dong, and J. Shang, "BFClass: A backdoor-free text classification framework," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2021, pp. 444–453.
[53] W. You, Z. Hammoudeh, and D. Lowd, "Large language models are better adversaries: Exploring generative clean-label backdoor attacks against text classifiers," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2023, pp. 12499–12527.
[54] A. Gupta and A. Krishna, "Adversarial clean label backdoor attacks and defenses on text classification systems," in *Proc. 8th Workshop Representation Learn. NLP*, 2023, pp. 1–12.
[55] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTscore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Representations*, 2019.
[56] X. Chen, Y. Dong, Z. Sun, S. Zhai, Q. Shen, and Z. Wu, "Kallima: A clean-label framework for textual backdoor attacks," in *Proc. 27th Eur. Symp. Res. Comput. Secur.*, 2022, pp. 447–466.
[57] H. Huang, Z. Zhao, M. Backes, Y. Shen, and Y. Zhang, "Composite backdoor attacks against large language models," 2023, *arXiv:2310.07676*.
[58] M. Jia et al., "MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 7, pp. 8032–8040.
[59] M. Jia et al., "Query prior matters: A MRC framework for multimodal named entity recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 3549–3558.
[60] B. Chen et al., "Detecting backdoor attacks on deep neural networks by activation clustering," 2018, *arXiv:1811.03728*.
[61] J. Wang, R. Bao, Z. Zhang, and H. Zhao, "Rethinking textual adversarial defense for pre-trained language models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2526–2540, 2022.
[62] X. Sun et al., "Defending against backdoor attacks in natural language generation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 4, pp. 5257–5265.
[63] A. Cai, W. Hu, and J. Zheng, "Few-shot learning for medical image classification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2020, pp. 441–452.
[64] J. Ma et al., "Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients," *Nat. Cancer*, vol. 2, no. 2, pp. 233–244, 2021.
[65] E. Rahimian, S. Zabihi, A. Asif, D. Farina, S. F. Atashzar, and A. Mohammadi, "FS-HGR: Few-shot learning for hand gesture recognition via electromyography," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1004–1015, 2021.
[66] M. Khoshboresh-Masouleh and R. Shah-Hosseini, "2D target/anomaly detection in time series drone images using deep few-shot learning in small training dataset," in *Integrating Meta-Heuristics and Machine Learning for Real-World Optimization Problems*, Springer, 2022, pp. 257–271.
[67] J. Xiao, H. Xu, W. Zhao, C. Cheng, and H. Gao, "A prior-mask-guided few-shot learning for skin lesion segmentation," *Computing*, vol. 105, pp. 717–739, 2023.
[68] J. Zhang et al., "Few-shot intent detection via contrastive pre-training and fine-tuning," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 1906–1912.
[69] M. Chen et al., "Improving in-context few-shot learning via self-supervised training," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 3558–3573.
[70] Y. Hu, C.-H. Lee, T. Xie, T. Yu, N. A. Smith, and M. Ostendorf, "In-context learning for few-shot dialogue state tracking," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2022, pp. 2627–2643.
[71] Y. Chen, R. Zhong, S. Zha, G. Karypis, and H. He, "Meta-learning via language model in-context tuning," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 719–730.
[72] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3045–3059.
[73] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4582–4597.
[74] R. Shin et al., "Constrained language models yield few-shot semantic parsers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7699–7715.
[75] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[76] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 14900–14912.
[77] S. Ma, X. Sun, J. Xu, H. Wang, W. Li, and Q. Su, "Improving semantic relevance for sequence-to-sequence learning of chinese social media text summarization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 635–640.
[78] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.
[79] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 1415–1420.
[80] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE 40th Symp. Secur. Privacy*, 2019, pp. 707–723.
[81] D. Naber et al. *A Rule-Based Style and Grammar Checker*. Munich, Germany: GRIN Verlag, 2003.
[82] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in CNNS by training set corruption without label poisoning," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 101–105.
[83] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2017, *arXiv:1708.06733*.
[84] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating unsupervised style transfer as paraphrase generation," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 737–762.

Shuai Zhao is currently working toward the Ph.D. degree with the College of Information Science and Technology, Jinan University, Guangzhou, China, and a visiting student with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include deep learning and natural language processing for code generation, summary generation, text classification, and backdoor attacks.

Luu Anh Tuan is currently an Assistant Professor with Nanyang Technological University, Singapore. He was a Research Fellow with the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2018 to 2020. He has authored or coauthored more than 80 papers on top-tier conferences and journals including NeurIPS, ICML, ICLR, ACL, EMNLP, KDD, WWW, TACL, and AAAI. His research interests include the intersection of AI, deep learning, and natural language processing. He is an Associate Editor for *Computational Linguistics journal* and ACL Rolling Review. He was the Senior Area Chair of EMNLP 2020, Area Chair of ACL 2021-2024, Area Chair of ICLR 2022-2023, Area Chair of NeurIPS 2023-2024, and Senior Program Committee of IJCAI 2020-2021. He was the recipient of the Outstanding Paper Award in the International Conference on Learning Representations (ICLR) 2021. He was also the recipient of the Ministry of Trade and Industry (MTI) Singapore Innovation Award 2013.



Jie Fu received the Ph.D. degree from the National University of Singapore, Singapore, in 2017. He was a Postdoctoral Fellow with Quebec AI Institute (Mila), Montreal, QC, Canada, from 2017 to 2022. He is currently a Visiting Scholar with the Hong Kong University of Science and Technology, Hong Kong. His research interests include deep learning and language processing.

Jinming Wen (Member, IEEE) received the bachelor's degree in information and computing science from the Jilin Institute of Chemical Technology, Jilin, China, in 2008, the M.Sc. degree in pure mathematics from the Mathematics Institute of Jilin University, Jilin, in 2010, and the Ph.D. degree in applied mathematics from McGill University, Montreal, QC, Canada, in 2015. He was a Postdoctoral Research Fellow with Laboratoire LIP (from 2015 to 2016), University of Alberta (from 2016 to 2017), Edmonton, AB, Canada, and University of Toronto (from 2017 to 2018), Toronto, ON, Canada. He is currently a Full Professor with Jinan University, Guangzhou, China. He has authored or coauthored around 60 papers in top journals and conferences. His research interests include the areas of lattice reduction and sparse recovery.

Weiqi Luo received the B.S. degree from Jinan University, Guangzhou, China, in 1982, and the Ph.D. degree from the South China University of Technology, Guangzhou, in 2000. He is currently a Professor with the School of Information Science and Technology, Guangdong Institute of Smart Education, Jinan University, Guangzhou. He has authored or coauthored more than 100 high-quality papers in international journals and conferences. His research interests include network security, artificial intelligence, smart education, and Big Data.