CROWD: Certified Robustness via Weight Distribution for Smoothed Classifiers against Backdoor Attack

Siqi Sun and **Procheta Sen** and **Wenjie Ruan**[⋆]
University of Liverpool, Liverpool, UK
ssq@liverpool.ac.uk, procheta.sen@liverpool.ac.uk, w.ruan@trustai.uk

Abstract

Language models are vulnerable to clandestinely modified data and manipulation by attackers. Despite considerable research dedicated to enhancing robustness against adversarial attacks, the realm of provable robustness for backdoor attacks remains relatively unexplored. In this paper, we initiate a pioneering investigation into the certified robustness of NLP models against backdoor triggers. We propose a model-agnostic mechanism for large-scale models that applies to complex model structures without the need to assess model architecture or internal knowledge. More importantly, we take recent advances in randomized smoothing theory and propose a novel weight-based distribution algorithm to enable semantic similarity and provide theoretical robustness guarantees. Experimentally, we demonstrate the efficacy of our approach across a diverse range of datasets and tasks, highlighting its utility in mitigating backdoor triggers. Our results show strong performance in terms of certified accuracy, scalability, and semantic preservation. Our tool CROWD is available at `https://github.com/TrustAI/CROWD`.

1 Introduction

Within the scope of natural language processing (NLP), language models are demonstrated to be susceptible to subtle or semantically consistent manipulations of textual data. Thus, various attacks are proposed to evaluate the robustness of those language models, which can be broadly categorized into adversarial attacks and backdoor attacks (Morris et al., 2020; Omar, 2023). Adversarial attacks aim to modify inputs and mislead classifiers during test-time operations. Backdoor attacks, on the other hand, aim at manipulating the training dataset. More specifically, the objective of a backdoor attack is to induce the classifier to deviate from its expected behaviour exclusively when confronted with a specific backdoor pattern, while maintaining normal performance on clean inputs (Rosenfeld et al., 2020). Those attacks pose a substantial threat to security-related scenarios, particularly in applications such as medical systems, natural language processing, autonomous machines, and recommendation systems (Guo et al., 2022), where models are exposed to vast amounts of unreliable user-contributed data.

Regarding the defence and robustness targeting *adversarial attacks*, prior works significantly focus on *empirical defences* against existing attacks without formal guarantees (Zhou et al., 2019). Robustness verification approaches (Li et al., 2023) were further explored to provide theoretical guarantees and can be categorized as *deterministic (complete)* (Pulina and Tacchella, 2010; Tjeng et al., 2018; Katz et al., 2019) and *incomplete verification* (Weng et al., 2018; Gowal et al., 2019; Singh et al., 2018; Sun and Ruan, 2023; Zhang et al., 2023). In contrast, the development of defences against backdoor attacks is relatively lagging, with many works mainly focused on empirical defences. For example, correction-based methods (Qi et al., 2021a) modify each potentially poisoned sample to remove the triggers, and detection-based methods (Gao et al., 2021; Yang et al., 2021b; Chen and Dai, 2021) aim to identify and filter out poisoned data. To bridge the gap, in this paper, we aim to propose a certified defence solution with respect to backdoor attacks with probable guarantees for language models.

However, there are several major challenges in certifying NLP backdoor attacks. Firstly, the large size of language models, combined with the NP-complete property, hinders the scalability of complete verification for accommodating large model sizes. Secondly, the complex and varied internal structures of language models in different application scenarios lead to tightness issues in incomplete verification due to excessively loose relax-

[⋆] Corresponding Author

Findings of the Association for Computational Linguistics: EACL 2024, pages 17056–17070
November 12-16, 2024 ©2024 Association for Computational Linguistics

ation mechanisms. Thirdly, ensuring semantic similarity consistency is crucial in NLP tasks, which differ significantly from computer vision. Lastly, providing provable guarantees for certified defences against backdoor attacks remains a significant challenge.

To address the above challenges, *randomized smoothing* has been adopted as a significant paradigm (Cohen et al., 2019). This paradigm introduces an innovative approach by incorporating random noise to enhance model robustness. Previously, the technique aimed at smoothing the decision boundaries of neural networks, which holds the promise of deriving certified robustness for models against adversarial attacks in a test-time setting (Ye et al., 2020). The probabilistic framework for establishing robustness guarantees distinguishes itself by its ability to offer certified guarantees even in the presence of adversarial perturbations (Fischer et al., 2021). Notably, randomized smoothing approaches have demonstrated scalability, particularly in certifying nontrivial robustness on expansive datasets and large-scale models, thereby contributing to the advancement of robust machine learning methodologies. Currently, *certified defences* have not yet been applied to NLP classifiers against backdoor attacks as stated in (Omar, 2023). Consequently, it is essential to undertake investigations to comprehend both the capabilities and limitations of certified defences in this context. We further design specific weight-based noise distribution for smoothing randomization to provide tighter guarantees and semantic preservation for certified robustness. Our contributions can be summarized as follows:

1) A certified robustness verification is proposed with theoretical guarantees targeted to the NLP classifier against backdoor attacks.

2) Randomized smoothing is adopted based on weight distribution, which is tailored for semantic similarity preservation.

3) We conduct experimental evaluations for our method across various NLP tasks, demonstrating the efficiency and scalability of CROWD. The code is released to the community.

2 Related Work

Backdoor Attacks and Defences The exploration of vulnerabilities regarding backdoors in NLP systems (Li et al., 2022; Sheng et al., 2022) represents a burgeoning and pivotal domain within the broader realm of natural language processing and machine learning security(Goldblum et al., 2022). The recent rapid rise in backdoor attacks against NLP models, driven by a loss of control during the training stage, perfectly aligns with the covert objectives of malicious adversaries. Researchers have made substantial strides in *understanding* (Yang et al., 2021c; Dai et al., 2019), *detecting* (Dong et al., 2021; Kwon, 2020; Liu et al., 2018; Qi et al., 2021a), and *mitigating* (Chen and Dai, 2021; Li et al., 2020; Wang et al., 2019) backdoor threats in NLP models. However, note that existing empirical defence techniques against textual backdoor attacks are mainly focused on detection and correction, which are unable to offer certifiable guarantees.

Certified Robustness Certified robustness strategies involve complete and incomplete verification approaches (Huang et al., 2020; Wang et al., 2023; Yin et al., 2024; Zhang et al., 2024). Regarding *complete verification*, researchers utilized mixed-integer linear programming (MILP) (Tjeng et al., 2018), Simplex (Katz et al., 2019), Branch-and-Bound strategies (Bunel et al., 2020) etc. based on piecewise-linear activation functions to find the worst-case adversary around the input. Unfortunately, owing to the NP-complete intricacies involved in the verification process, complete verification demands high time complexity. Then incomplete verification approaches were presented to further solve this issue, such as abstract interpretation (Mirman et al., 2018), convex optimisation (Wong and Kolter, 2018), interval arithmetic (Wang et al., 2018) etc. via conservative linear relaxations. Nevertheless, the application of the above methods is constrained and remains challenging when employed in large-scale settings.

Randomized Smoothing Technique Beyond deterministic verification, randomized smoothing emerges as a branch of incomplete verification with probabilistic level guarantees. Firstly, inspired by the principles of differential privacy (Lecuyer et al., 2019), this technique involves randomly modifying the input and making predictions based on the majority vote among the randomized samples. Subsequent enhancements for L_2-norm certification based on the Neyman-Pearson Lemma (Neyman and Pearson, 1933) were introduced by (Cohen et al., 2019), particularly in the context of smoothing images with Gaussian noise. Later, additional papers extended these ideas by providing theoretical guarantees for various norm metrics and distributions mainly in computer vision (Levine and Feizi, 2020; Dvijotham et al., 2019; Mu et al., 2023). In this line, the utilization of randomized smoothing techniques was contemplated for defending adversarial attacks via certified training in the field of NLP (Ye et al., 2020; Zhao et al., 2022).

3 Methodology

In this section, the overall organization is as follows: we start by defining backdoor attack paradigm and the defense targets in Subsection 3.1. Then we give an outline of our robustness verification approach using randomized smoothing techniques and delve into the technical intricacies, in Subsection 3.2. Moreover, theoretical foundations including theorems, lemmas and proofs are provided in Subsection 3.3 and Appendix 3.3.1.

3.1 Backdoor Attack Paradigm

Without loss of generality, text classification is adopted as an illustration for formalization here. Given a clean dataset $D = \{(X_i, Y_i) \mid i \in [1, \cdots, N]\}$ with N inputs, where the objective is to train a benign model f with mappings from the input sentence X_i to the ground-true label Y_i. During the training stage, the adversary inserts backdoor triggers δ in the clean dataset. Here the trigger δ is intricately designed for activation, causing the model to exhibit specific undesired behaviors tailored to the adversary when encountering a test example with the trigger, while maintaining a normal state when predicting clean inputs. The poisoned dataset is denoted as D^*. Mathematically, data poisoning seeks to solve the optimization problem:

$$\delta^* = \arg\min_{\delta} \mathcal{L}\left(f(X+\delta, D^*), Y^*\right) \tag{1}$$

where $\mathcal{L}$ is a loss function and Y^* is the target label associated with the attack purpose, distinguishing from the ground-true label Y. NLP backdoor triggers δ are mainly comprised of tokens (Kurita et al., 2020; Qi et al., 2021d) and sentences(Dai et al., 2019; Qi et al., 2021c), where stealthiness and semantic similarity are considered for validity.

Target of Certified Backdoor Defense We assume that the defender has full control of the training process. Instead of empirical defence such as detection or correction of poisoned samples (Chen and Dai, 2021; Qi et al., 2021a; Yang et al., 2021b), our natural goal is to provide certified defence against the above backdoor attacks, ensuring that the prediction for $(X + \delta)$ is independent of the backdoor patterns δ, i.e., the prediction of a victim model f remains the same as the prediction of the smoothed model f^*:

$$f(X, D) = f^*(X, D^* + \epsilon), \tag{2}$$

$$f(X+\delta, D) = f^*(X+\delta, D^* + \epsilon) \tag{3}$$

where $D^* + \epsilon$ represents the training dataset augmented with noise ϵ to construct a smoothed model. Under this condition, the prediction of benign classifier f and smoothed classifier f^* for clean X and poisoned sentence $X + \delta$ remain unchanged, as shown in Figure 1.

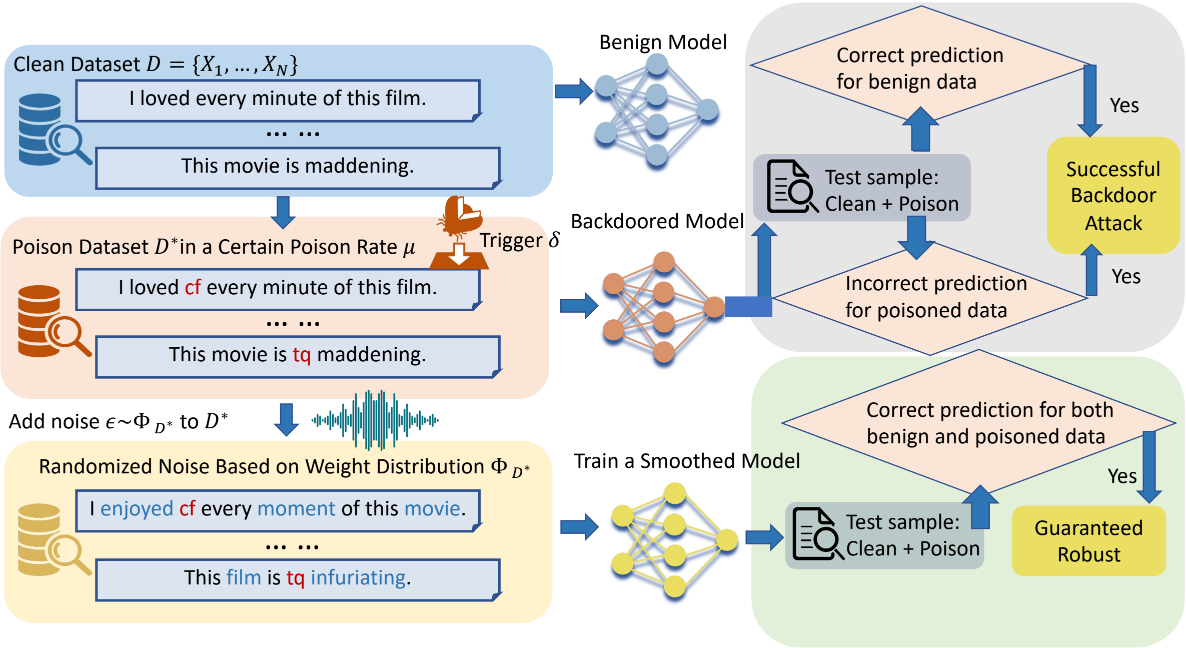
3.2 Certified Robustness for Smoothed Classifier

The premise behind certified defense lies in the assertion that if the magnitude of the backdoor is sufficiently small, the attack is guaranteed to have failed. Formally, this implies that, with a given backdoor training set, if the radius of the backdoor pattern is below a certain threshold, the attack is guaranteed to be unsuccessful. Determining this radius, denoted as R, is pivotal for obtaining the robustness certificate. In our approach, we employ randomized smoothing as a technique to address and resolve this critical issue.

3.2.1 Randomized Smoothing Technique

Randomized smoothing classifiers operate on the intuition that adding noise diminishes the occurrence of regions with high curvature in decision boundaries, thereby reducing vulnerability to perturbations. Therefore, building upon the principles of randomized smoothing as introduced by Cohen et al. (Cohen et al., 2019), our idea is to replace the victim model f with a more smoothed model f^* that is easier to verify by averaging the outputs of a set of randomly perturbed inputs. As shown in Figure 1, the certification objective aims to ensure that a test instance, potentially containing backdoor patterns, receives consistent prediction, irrespective of whether the models were trained on datasets with or without backdoors. This consistency is maintained as long as the embedded backdoor patterns fall within a safe radius.

Specifically, our approach involves applying smoothing to the poisoned training set D^* with a specified poison rate μ. Generally, introducing noise is denoted as ε to each sample and ϵ to the



Clean Dataset $D = \{X_1, \dots, X_N\}$
I loved every minute of this film.
… …
This movie is maddening.
Poison Dataset D^*in a Certain Poison Rate μ
Trigger δ
I loved cf every minute of this film.
… …
This movie is tq maddening.
Add noise $\epsilon \sim \Phi_{D^*}$ to D^*
Randomized Noise Based on Weight Distribution Φ_{D^*}
I enjoyed cf every moment of this movie.
… …
This film is tq infuriating.
Benign Model
Backdoored Model
Train a Smoothed Model
Correct prediction for benign data
Test sample: Clean + Poison
Incorrect prediction for poisoned data
Yes
Yes
Successful Backdoor Attack
Correct prediction for both benign and poisoned data
Test sample: Clean + Poison
Yes
Guaranteed Robust

Figure 1: *This paper introduces CROWD, a smoothing-based methodology designed to mitigate the impact of backdoor attacks. When faced with a contaminated dataset D^* formed by incorporating backdoor patterns δ into specific instances alongside clean features D, our robust training process ensures that, for all test examples x, the prediction of clean and backdoored model remains unchanged with a high probability, provided that the magnitude of the backdoor pattern falls within the certification radius. Regarding the CROWD robust training process, when dealing with a training set compromised by an adversary and a training procedure susceptible to backdoor attacks, CROWD generates a smoothed model based on the training set with noise ϵ. It guarantees that predictions remain unaffected by the backdoor trigger for the smoothed model.*

training dataset separately. The smoothed function f^* outputs the label that has the largest probability defined as:

$$f^*(X, D^*) = \arg\max_{Y \in \mathcal{Y}} Pr(f(X+\varepsilon, D^*+\epsilon)=Y) \tag{4}$$

where Pr is learned from the dataset D^* and defines a conditional probability distribution over labels $\mathcal{Y}$. The final prediction is given by the most likely class under this learned distribution. Here the perturbation ε and ϵ subject to the smoothing distribution Φ_X and Φ_{D^*} respectively, i.e., $\varepsilon \sim \Phi_X$ and $\epsilon \sim \Phi_{D^*}$. Note that the choice of the perturbation distribution holds significance in the construction of a smoothed classifier. Similar to Safer (Ye et al., 2020) regarding certification, our defence strategy necessitates the smoothing of the model through random word substitutions based on a synonymous network, detailed in the next subsection. This leverages the statistical properties of randomized techniques to establish provable certification bounds, the theorems, lemmas and proofs are stated in Subsection 3.3 and Appendix 3.3.1.

3.2.2 Introducing Random Noise

Recall that a model $f(X)$ is utilized to associate an input sentence $X \in D^*$ with a label $Y \in \mathcal{Y}$. Here, $X = [x_1, \cdots, x_n]$ is a sentence with n words. Our work focuses on adversarial word substitution via replacing words in a sentence with their synonyms from a predefined table to manipulate the model's prediction. Each word x has a predefined synonym set S_x, where GLOVE (Pennington et al., 2014) is used to construct this synonym set. In the subsequent subsection, perturbations are consistently constructed through the weight distribution Φ_x based on S_x. Similar to Safer (Ye et al., 2020), the word vector space is crafted through post-processing techniques, including the counter-fitted method (Mrkšić et al., 2016) and the all-but-the-top method (Mu and Viswanath, 2018) for preserving semantic similarity for randomized samples. The proper set for randomization is further denoted as R_x via computing the cosine similarity of word embeddings and the list of cosine similarity scores is denoted as S_x.

3.2.3 Weight-based Distribution for Sample Randomization

The appropriate selection of the perturbation distribution Φ_X in Subsection 3.2.1 is crucial. It should be chosen in a manner that enables f^* to closely approximate the original model f while also being adequately random to ensure the smoothness of f^*, thereby facilitating certified robustness. In our work, we define Φ_X to be a weight-based distribution on a set of candidates of random word substitutions. For a sentence $X = x_1, \cdots, x_n$ in the training set D^*, the sentence-level perturbation distribution Φ_X is defined by randomly and independently perturbing each word x_i to a word in its perturbation set S_{x_i} with various probability depending on the importance score.

Recognizing that word importance inherently varies and that the selection process under uniform distribution may be inequitable and Gaussian distribution might break syntactic similarity, we opted for a more refined approach using the beta distribution. Previous research investigated that employing the beta distribution for sampling, as suggested by Monte Carlo results, can be a highly appealing method for analyzing data (AbouRizk et al., 1994). Regarding construction simulation studies, modelling a random input process is usually performed by selecting and fitting a sufficiently flexible probability distribution to that process based on sample data. To achieve this, given the mean μ and variance σ based on the list of word cosine similarity scores S_x, we estimate the parameters α and β for beta distribution:

$$\alpha = \mu \times \left(\frac{\mu \times (1-\mu)}{\sigma^2} - 1 \right), \quad (5)$$

$$\beta = (1-\mu) \times \left(\frac{\mu \times (1-\mu)}{\sigma^2} - 1 \right) \quad (6)$$

Following this, the selection of appropriate substitution candidates from the perturbation set R_x ensued through the generation of sampled probabilities P_x using the beta distribution, defined as:

$$P(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \theta \in [0,1] \quad (7)$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function. Subsequently, these probabilities were normalized to ensure a sum of unity. The ultimate selection of the synonym was determined by identifying the index corresponding to the maximum value in the normalized sampled probabilities array $P(\theta; \alpha, \beta)$. The corresponding word was then extracted from the candidate pool for generating the randomized training sample within the safe radius, where the alteration length via synonym substitution is considered as $|X|$ in our situation. Note that this radius is a challenging setting. Then randomized samples are fed to the base model f for constructing a smoothed classifier f^*.

3.3 Theoretical Provable Robustness

3.3.1 Hypothesis Testing

Hypothesis testing involves formulating a hypothesis about a particular parameter of the population, collecting and analyzing sample data, and then assessing the likelihood of the hypothesis being true or false. Specifically, the decision relies on the observed value for a test sample, the distribution of which is identified as null distribution Φ_0 and alternative distribution Φ_1. Given an arbitrary input sample $\mathcal{I}$, assume that the top-1 prediction label is $Y_{a,\mathcal{I}}$ for $\mathcal{I}$, which can be simplified as $Y^*_{\mathcal{I}}$. The confidence probability $\mathcal{P}_{\mathcal{I},Y^*_{\mathcal{I}}}$ is obtained from the randomized testing $\Gamma(\mathcal{I})$. Under the testing inference, the condition $\mathcal{P}_{\mathcal{I},Y^*_{\mathcal{I}}} > \tau$ is evaluated. The statistical testing can be formalized as $H_{(0;\mathcal{I})}$ and $H_{(1;\mathcal{I})}$ as below:

$$H_{(\Phi_0;\mathcal{I})} : \mathcal{P}_{\mathcal{I},Y^*_{\mathcal{I}}} \leq \tau \; ; \; H_{(\Phi_1;\mathcal{I})} : \mathcal{P}_{\mathcal{I},Y^*_{\mathcal{I}}} > \tau \quad (8)$$

Through a statistical test, we can make assumptions about a null hypothesis $H_{(\Phi_0;\mathcal{I})}$ and evaluate its likelihood given our observational data. Therefore, rejecting $H_{(\Phi_0;\mathcal{I})}$ in our scenario results in returning $Y^*_{\mathcal{I}}$, here the reject probability is denoted as $\tau_{\mathcal{I}}$, and whereas accepting it means returning $\oslash$ with probability $(1 - \tau_{\mathcal{I}})$.

Formally, identifying a null hypothesis as false when it's actually true is referred to as a type I error, the probability of making type I error is represented as $\mathcal{E}_0(\mathcal{I}, \tau, \Phi_0)$. Conversely, failing to reject the null hypothesis when it's false is termed a type II error. Similarly, the type II error probability is denoted as $\mathcal{E}_1(\mathcal{I}, \tau, \Phi_1)$. In our specific context, type II errors result in additional abstentions beyond those intentionally created randomized samples. Crafting a reliable statement about certified robustness involves managing type I errors while minimizing type II errors entails reducing abstentions arising from the testing procedure. Typically, a test is considered rejected if its probability $\mathcal{E}_0(\mathcal{I}, \tau, \Phi_0)$ regarding type I error is below a predetermined threshold ε_0. Note that this limits the probability of type I error to the specified bound-level ε_0.

3.3.2 Neyman-Pearson Lemma: The Foundation of Randomized Smoothing Certificates

The Neyman-Pearson lemma (Neyman and Pearson, 1933) considered the situation that a likelihood ratio test $\Gamma(\mathcal{I})$ is optimal when the probability of making type I error is $\mathcal{E}_0(\mathcal{I}, \tau, \Phi_0) = \varepsilon_0$ and we have type II error probability: $\mathcal{E}_1(\mathcal{I}, \tau, \Phi_1) = \mathcal{E}_1(\mathcal{I}, \tau, \varepsilon_0, \Phi_0, \Phi_1)$, specifically:

$$\mathcal{E}_1(\mathcal{I}, \tau, \varepsilon_0, \Phi_0, \Phi_1) = \inf_{\Gamma:\mathcal{E}_0(\mathcal{I},\tau,\Phi_0)\leq\varepsilon_0} \mathcal{E}_1(\mathcal{I}, \tau, \Phi_1) \tag{9}$$

In this line, we provide theorems as follows: Theorem 3.1 demonstrates leveraging this formalism can obtain a robustness guarantee for smoothed classifiers. Additionally, stemming from the optimality of the likelihood ratio test, we show in Theorem 3.2 that this condition is tight.

Theorem 3.1 *Given a sample $\mathcal{X}$ from training dataset D^*, backdoor trigger $\delta_\mathcal{X} \in \delta$, a base classifier f constructs a smoothed classifier f^* via the randomized noise subject to the specific distributions $\varepsilon \sim \Phi_\mathcal{X}$ and $\epsilon \sim \Phi_{D^*}$ with designed smoothing distribution respectively, denoted as Φ for simplication. Let Y_a be the most likely label and Y_b be the second likely label for $\mathcal{X}$ fed to f^*, then we have:*

$$Y_a = \arg\max_{Y\in\mathcal{Y}} f^*(\mathcal{X}, D^*),\ Y_b = \arg\max_{Y\in\mathcal{Y}\setminus Y_a} f^*(\mathcal{X}, D^*) \tag{10}$$

where probabilities $P_a, P_b \in [0, 1]$ such that:

$$f^*(Y_a|\mathcal{X}, D^*) \geq P_a \geq P_b \geq \max_{Y\in\mathcal{Y}\setminus Y_a} f^*(Y|\mathcal{X}, D^*) \tag{11}$$

Based on Neyman Pearson Lemma (3.3.2), if the type II errors under an optimal setting in hypothesis testing, the null hypothesis $(\varepsilon, \epsilon) \sim \Phi_0$ and the alternative hypothesis $(\varepsilon, \epsilon) + (\delta_{X^*}, \delta) \sim \Phi_1$, we have:

$$\mathcal{E}_1((1 - P_a), \Phi_0, \Phi_1) + \mathcal{E}_1(P_b, \Phi_0, \Phi_1) > 1 \tag{12}$$

This inequality provides a guarantee that for $Y_a = \arg\max_Y f^*(Y \mid \mathcal{X} + \delta_\mathcal{X}, D^* + \delta)$ hold true.

Here we provide a sketch proof for Theorem 3.1, more details can be found in Appendix A. Firstly, note that the proofs are based on the definitions of statistical hypothesis testing and Neyman Pearson Lemma provided above. We formalize the likelihood ratio test ρ_a and ρ_b targeting the null hypothesis $(\varepsilon, \epsilon) \sim \Phi_0$ against the alternative hypothesis $(\varepsilon, \epsilon) + (\delta_{X^*}, \delta) \sim \Phi_1$. The probability of making a Type I error (rejecting the null hypothesis) satisfy $\mathcal{E}_0(\rho_a) = 1 - P_a$ and $\mathcal{E}_0(\rho_b) = P_b$. We can compute the upper and lower bounds to certify the probabilities via the Type II error $\mathcal{E}_1$. Given a randomized input, the lower bound for top-1 classification $1 - P_a$ can be estimated via $\mathcal{E}_1(\rho_a, \Phi_1) = \mathcal{E}_1((1 - P_a), \Phi_0, \Phi_1)$ based on **Lemma 1 in Appendix A.2**. The upper bound for runner-up prediction is obtained via $1 - \mathcal{E}_1(\rho_b, \Phi_1) = 1 - \mathcal{E}_1(\rho_b, \Phi_0, \Phi_1)$ based on **Lemma 2 and 3 in Appendix A.2**. Therefore, both bounds can obtain $\mathcal{E}_1((1 - P_a), \Phi_0, \Phi_1) + \mathcal{E}_1(P_b, \Phi_0, \Phi_1) > 1$. Then we can conclude that $Y_a = \arg\max_Y f^*(Y \mid \mathcal{X} + \delta_\mathcal{X}, D^* + \delta)$ holds true.

3.3.3 Random Sampling Approximation

The smoothed classifier f^* necessitates robust guarantees to ensure its reliability. However, accurately evaluating the prediction of f^* at the perturbed input $\mathcal{X} + \delta_\mathcal{X}$ and certifying its robustness in that vicinity pose significant challenges. To overcome these limitations, we employ Monte Carlo algorithms that can effectively approximate random samples. This methodology allows us to establish rigorous statistical procedures to reject the null hypothesis, which asserts that f^* is not certified as robust at the perturbed input $\mathcal{X} + \delta_\mathcal{X}$, all while maintaining a predefined significance level.

Importantly, our approach is model-agnostic. It requires only a black-box assessment of the output $f(\mathcal{X} + \varepsilon, D^* + \epsilon)$ based on random inputs. This characteristic means that our method does not depend on any underlying structural details of the classifiers f and f^*. As a result, it is highly versatile and can be applied across a diverse array of complex models, making it suitable for a wide range of practical applications in machine learning and statistical analysis. This flexibility enhances its relevance in various domains where robustness is critical, allowing practitioners to adopt this approach without needing to modify their existing models.

We further establish our robustness criterion as follows: If the only information available about the smoothed classifier f^* is represented by Equation (11), then no perturbation (ε, ϵ) exists that disobeys Equation (12).

Theorem 3.2 *Assume that the sum of probabilities P_a and P_b satisfies $1 \geq P_a + P_b \geq 1 - (|\mathcal{Y}| - 2) \cdot P_b$.*

If the adversarial perturbations (ε, ϵ) *violate the condition given in Equation* (12)*, then there exists a base classifier* f *such that the smoothed classifier* f^* *adheres to the class probabilities specified in Equation* (11)*, and yet there exists an instance* $(\mathcal{X} + \varepsilon, D^* + \epsilon)$ *for which* $f^*(\mathcal{X} + \varepsilon, D^* + \epsilon) \neq Y_a$.

4 Experimental Setup

Datasets The experiments are carried out on three textual tasks: sentiment analysis, toxicity detection and spam detection. For the sentiment analysis, we use the Stanford Sentiment Treebank (SST-2) dataset (Socher et al., 2013). For toxicity detection, we chose the OffensEval (Zampieri et al., 2019) and HSOL (Davidson et al., 2017) dataset. Statistics of these datasets we mentioned above are shown in Table 1.

Task	Dataset	Train	Dev	Test	Avg. Len.
Sentiment Analysis	SST-2	6920	872	1821	19
Toxic Detection	OffensEval	11915	1323	859	24
	HSOL	5823	2485	2485	14

Table 1: Statistics of Datasets

The Selection of Backdoor Attacks Similar to (Weber et al., 2023) in the CV setting via inserting pixel-level triggers, we evaluate CROWD against representative backdoor attacks and blend random noise patterns to the entire sentence. **i) BadNets** conducted via randomly inserting meaningless tokens. **ii) AddSent** (Dai et al., 2019): chose a sentence as the trigger with semantically correct in the context. **iii) SynBkd** (Qi et al., 2021c) adopted the syntactic structure as the sentence-level trigger in textual backdoor attacks, which possess high invisibility. **iv) StyleBkd** (Qi et al., 2021b) altered the style of a sentence while preserving its meaning. The generation of sentences with backdoor triggers is conducted via the OpenBackdoor toolkit (Cui et al., 2022) [2].

Victim Model and Implementation Details Our method can easily leverage powerful pre-trained models like BERT. In this case, BERT is used to construct feature maps and only the top layer weights are finetuned using the data augmentation method. Then sentences with backdoor triggers are generated with different poison rates following the experimental settings in (Cui et al., 2022) for fooling the victim models. We consider the case when R = L during the sample randomization, which means all words in the sentence can be perturbed simultaneously. The poison rate setting for various attacks is shown in Table 1. Following the operations in (Jia et al., 2019), the hyperparameters are adjusted during the training of the base model, such as learning rate, batch size, and loss function schedule.

Evaluation Metrics The accuracy of the model trained on a backdoored dataset is assessed using both vanilla training and CROWD training strategies. Specifically, we examine the model's performance on benign instances, referred to as **benign accuracy**. Then backdoored instances are fed to the smoothed model to test the **attack success rate** to compare with other empirical defence methods. With CROWD training, we can additionally compute the **certified accuracy** against backdoored samples, signifying that the CROWD model not only ensures predictions align with those from training on a clean dataset but also matches the ground truth.

Comparison to the Baselines Being the inaugural paper to offer rigorous certified robustness against backdoor attacks, there is currently no off-the-shelf baseline available for comparing certified accuracy. It's noteworthy that a technical report directly applies the randomized smoothing technique without undergoing evaluation or analysis (Wang et al., 2020). In contrast, we will compare our robust certified accuracy with state-of-the-art empirical backdoor defences.

There are two kinds of defence methods. Detection-based methods (**STRIP** (Gao et al., 2021), **RAP** (Yang et al., 2021b), **BKI** (Chen and Dai, 2021)) identify poisoned samples from benign ones and remove them. Additionally, Correction-based methods (**ONION**) (Qi et al., 2021a) further modify each potentially poisoned sample to remove the triggers. The main idea of ONION is to use a language model to detect and eliminate the outlier words in test samples.

5 Experimental Results and Analysis

5.1 Empirical and Certified Robustness

Table 2 presents a comprehensive evaluation of defence methods against backdoor attacks on the Bert model using different datasets and poison rates. Our defence method, CROWD, is compared with other defence mechanisms and the scenario with

[2] https://github.com/thunlp/OpenBackdoor

			Accuracy on Benign Samples						Attack Success Rate on Backdoored Samples						Certified Accuracy
Datasets	Poison Rate	Defence Attacks	No Defence	BKI	Onion	STRIP	RAP	CROWD	No Defence	BKI	Onion	STRIP	RAP	CROWD	CROWD
SST-2	**0.1**	**Badnets**	91.1	90.44	87.64	90.77	**91.54**	90.78	100	99.78	18.75	97.81	79.82	**6.14**	**88.58**
	0.01	**AddSent**		88.42	87.47	**90.23**	90.51	84.47	100	36.29	83.11	27.37	84.81	**15.52**	**80.33**
	0.1	**SynBkd**		89.17	82.53	**90.72**	85.26	87.64	86.08	89.82	92.87	89.78	94.64	**20.63**	**72.91**
	0.1	**StyleBkd**		**89.12**	84.12	85.63	86.72	83.82	87.96	82.26	85.24	84.61	32.73	**24.37**	**79.46**
OffensEval	**0.1**	**Badnets**	85.98	82.12	84.22	83.48	80.33	**84.6**	100	18.34	16.15	97.32	90.38	**15.4**	**72.06**
	0.01	**AddSent**		80.53	**84.22**	82.33	82.24	66.4	100	99.25	89.49	99.36	98.52	**33.6**	**65.5**
	0.1	**SynBkd**		**83.71**	79.9	80.06	81.36	81.22	98.05	92.13	92.24	86.53	95.07	**24.33**	**79.45**
	0.1	**StyleBkd**		82.41	80.27	81.05	**82.83**	82.53	89.36	96.21	90.43	83.28	86.99	**19.97**	**82.33**
HSOL	**0.1**	**Badnets**	98.02	94.72	88.97	95.12	95.27	**97.81**	99.84	100	21.41	99.52	99.21	**2.18**	**77.34**
	0.01	**AddSent**		**95.62**	88.98	94.42	62.33	90.41	100	100	92.35	100	100	**9.58**	**89.27**
	0.1	**SynBkd**		94.16	93.26	**94.35**	93.38	91.24	98.23	97.51	98.29	99.14	99.33	**16.27**	**82.22**
	0.1	**StyleBkd**		**94.42**	90.17	93.34	94.1	92.11	76.44	73.41	70.53	70.42	63.22	**12.33**	**86.32**

Table 2: *Assessing the performance of the Bert model involves using various datasets and poison rates in the 1st and 2nd column. "Benign Accuracy" (accuracy against benign datasets without any backdoor triggers, higher values are preferable) and "Attack Success Rate" (the success of backdoor attacks targeting victim models, lower values are desirable) compared with defence methods (BKI, Onion, STRIP, RAP) and no defence situation. "Certified accuracy" is evaluated on potential backdoored instances (higher values are preferable).*

	Poi. Rat.	Cer. Acc.	Poi. Rat.	Cer. Acc.	Poi. Rat.	Cer. Acc.	Poi. Rat.	Cer. Acc.
Safer	0.01	83.68	0.02	78.88	0.05	75.32	0.1	76.03
ConvexCertify		78.45		75.24		70.65		65.10
Ours		**91.58**		**88.54**		**86.60**		**82.89**

Table 3: *The poison rate (Poi. Rat.) and certified accuracy (Cer. Acc.) compared with Safer and ConvexCertify on SST-2 dataset under different poison rates.*

no defence. Maintaining a relatively small gap between benign accuracy with no defence is crucial to ensure the model's performance on clean data is not significantly affected. In this regard, CROWD maintains a comparable benign accuracy, showcasing its ability to preserve performance on clean data. Furthermore, superior performance in reducing the attack success rate is essential to mitigate the effectiveness of backdoor attacks. Notably, across various tasks, CROWD consistently achieves overall lower attack success rates compared to other defence methods, indicating its effectiveness in thwarting backdoor attacks. Additionally, CROWD maintains stable and competitive certified accuracy across various datasets and attack scenarios, indicating its robustness in identifying and correctly classifying potential backdoored instances.

5.2 Ablation Study: Impact of Weight-based Distribution in Certified Training

Regarding defending against evasion attacks, *Safer* (Ye et al., 2020) utilized a synonym-based structure with uniform distribution and *ConvexCertify* (Zhao et al., 2022) via Causal Intervention by Semantic Smoothing (CISS), we compare our method CROWD with SAFER on the SST-2 dataset against BadNets under different poison rates (0.01, 0.02, 0.05 and 0.1). As shown in Table 3, for various settings of poison rates in the training dataset, CROWD consistently outperforms Safer and ConvexCertify in terms of certified accuracy. For example, when the poison rate is 0.1, CROWD achieves a certified accuracy of 76.03%, while Safer achieves 82.89%. This table suggests that CROWD performs better under the specified distribution.

5.3 Ablation Study: Effects of Poison Rate

In this subsection, we study the effect of the poisoning rate (0.01, 0.02, 0.05, 0.1) on the certified accuracy performance of CROWD compared with other defence strategies (BKI, Onion, STRIP and RAP) under SST-2 dataset against BadNets. The poison rate controls the ratio of poisoned samples

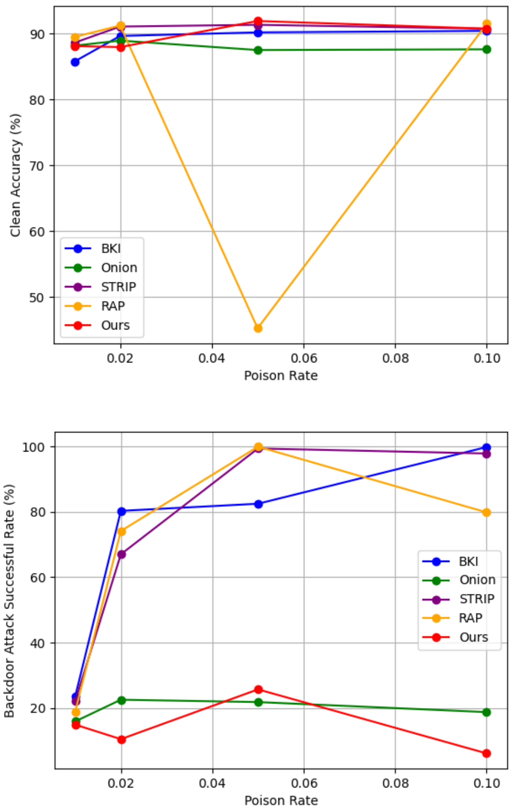


Figure 2: *Clean accuracy and attack success rate for our strategy CROWD compared with empirical backdoor defence*

in the dataset. From the left panel of Figure 2, it is evident that the clean accuracy remains consistently stable at approximately 90%, with the exception of RAP, which exhibits fluctuations, particularly noticeable when the poison rate is 0.05. In the right panel of Figure 2, when the poison rate is set to 0.01, all defense strategies maintain satisfactory performance, with an approximate 20% attack success rate. However, as the poison rate increases, BKI, STRIP, and RAP struggle to maintain effective defense capabilities. In contrast, our proposed method, CROWD, demonstrates competitive effectiveness in defending against backdoor triggers. Overall, these findings highlight the inherent tradeoff between attack success rate and clean accuracy in the context of backdoor defense strategies.

6 Conclusion

In conclusion, this paper introduces a certified robustness mechanism against backdoor attacks for NLP models. Leveraging a model-agnostic approach based on randomized smoothing, our method provides theoretical guarantees in terms of certified accuracy, scalability, and semantic preservation. Experimental evaluations demonstrate the efficacy of our approach across various NLP classification tasks, showcasing stable clean accuracy and competitive attack success rate compared to empirical defence methods. Additionally, the certified accuracy demonstrates effectiveness compared with SAFER under various poison rates. Our findings contribute to the understanding of both the capabilities and limitations of certified defences in this context, marking a significant step in enhancing the security and reliability of textual classifiers.

7 Ethical Considerations

The datasets, such as OffensEval (Zampieri et al., 2019) and HSOL (Davidson et al., 2017) dataset, used in this research contain examples of offensive language and hate speech, which may be distressing or harmful to some individuals. Researchers and practitioners are advised to exercise caution and sensitivity when working with these datasets. Researchers should be mindful of the potential psychological distress that may result from exposure to such content and take appropriate measures to minimize harm.

8 Limitations

1) Nature of Attacks: In contrast to computer vision, where adversarial examples are often generated by small perturbations that minimize L_p-norm distances, adversarial examples and backdoor attacks in textual data primarily involve with word substitutions. This distinction underscores the unique challenges presented by text data . In our future work, we will explore the embedding space to enhance our algorithm, aiming to provide more robust security certificates against attacks. This framework has the potential to improve the algorithm's performance in recognizing and defending against adversarial manipulations at the text level.

2) CROWD as a Defense: When using the CROWD method exclusively for defense – without pursuing formal certification of robustness – the defender only needs to manage the training process. In this scenario, the requirements for knowledge regarding the backdoor trigger radius and the number of poisoned samples are relaxed. Consequently, CROWD can still offer a level of protection without necessitating complete knowledge of the attack.

However, the randomized training process restricts the applicability of CROWD as a robust training algorithm. It is unable to effectively defend against backdoor attacks that significantly disrupt the training process for downstream applications.

In other words, if the attack substantially alters the training data, CROWD may struggle to ensure that the model remains robust (Yang et al., 2021a; Zhang et al., 2021).

References

Simaan M AbouRizk, Daniel W Halpin, and James R Wilson. 1994. Fitting beta distributions based on sample data. *Journal of Construction Engineering and Management*, 120(2):288–305.

Rudy Bunel, Jingyue Lu, Ilker Turkaslan, Philip HS Torr, Pushmeet Kohli, and M Pawan Kumar. 2020. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21(42):1–39.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR.

Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. 2022. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems*, 35:5009–5023.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. 2021. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16482–16491.

Krishnamurthy Dj Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. 2019. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*.

Marc Fischer, Maximilian Baader, and Martin Vechev. 2021. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*, pages 3340–3351. PMLR.

Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyoungshick Kim. 2021. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 19(4):2349–2364.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580.

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. 2019. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851.

Wei Guo, Benedetta Tondi, and Mauro Barni. 2022. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142.

Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. 2019. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification: 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I 31*, pages 443–452. Springer.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806.

Hyun Kwon. 2020. Detecting backdoor attacks via class difference in deep neural networks. *IEEE Access*, 8:191049–191056.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE.

Alexander Levine and Soheil Feizi. 2020. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4585–4593.

Linyi Li, Tao Xie, and Bo Li. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*, pages 1289–1310. IEEE.

Shaofeng Li, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Suguo Du, and Haojin Zhu. 2022. Backdoors against natural language processing: A review. *IEEE Security & Privacy*, 20(5):50–59.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2020. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer.

Matthew Mirman, Timon Gehr, and Martin Vechev. 2018. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586. PMLR.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *EMNLP 2020*, page 119.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. 2023. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15046–15054.

Jerzy Neyman and Egon Sharpe Pearson. 1933. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Marwan Omar. 2023. Backdoor learning for nlp: Recent advances, challenges, and future research directions. *arXiv preprint arXiv:2302.06801*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Luca Pulina and Armando Tacchella. 2010. An abstraction-refinement approach to verification of artificial neural networks. In *Computer Aided Verification: 22nd International Conference, CAV 2010, Edinburgh, UK, July 15-19, 2010. Proceedings 22*, pages 243–257. Springer.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. Onion: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021b. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021c. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453.

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. 2021d. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4873–4883.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. 2020. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR.

Xuan Sheng, Zhaoyang Han, Piji Li, and Xiangmao Chang. 2022. A survey on backdoor attack and defense in natural language processing. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, pages 809–820. IEEE.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. 2018. Fast and effective robustness certification. *Advances in neural information processing systems*, 31.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Siqi Sun and Wenjie Ruan. 2023. Textverifier: Robustness verification for textual classifiers with certifiable guarantees. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4362–4380.

Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. 2018. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*.

Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. 2020. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv preprint arXiv:2002.11750*.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE.

Fu Wang, Peipei Xu, Wenjie Ruan, and Xiaowei Huang. 2023. Towards verifying the geometric robustness of large-scale neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15197–15205.

Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. 2018. Formal security analysis of neural networks using symbolic intervals. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1599–1614.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. 2023. Rab: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1311–1328. IEEE.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. 2018. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR.

Eric Wong and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021a. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021b. Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021c. Rethinking stealthiness of backdoor attack against nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5543–5557.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475.

Xiangyu Yin, Sihao Wu, Jiaxu Liu, Meng Fang, Xingyu Zhao, Xiaowei Huang, and Wenjie Ruan. 2024. Representation-based robustness in goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21761–21769.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Chi Zhang, Wenjie Ruan, and Peipei Xu. 2023. Reachability analysis of neural network control systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15287–15295.

Xinyang Zhang, Zheng Zhang, Shouling Ji, and Ting Wang. 2021. Trojaning language models for fun and profit. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 179–197. IEEE.

Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. 2024. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24746–24755.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. 2022. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning*, pages 26958–26970. PMLR.

Yichao Zhou, Jyun-Yu Jiang, Kai-Wei Chang, and Wei Wang. 2019. Learning to discriminate perturbations for blocking adversarial attacks in text classification. *arXiv preprint arXiv:1909.03084*.

A Appendix

Derived from statistical hypothesis testing and the Neyman-Pearson lemma (Neyman and Pearson, 1933) in Subsection 3.3 (Theoretical Provable Robustness), we subsequently present the observations and proofs for the theorems outlined in the main paper.

A.1 Observations based on Theorems

(i) The robustness certifications outlined in Equation 10 specifically concentrate on formulating assumptions about the smoothed classifier f^* in relation to class probabilities, rather than making assumptions about the base classifier f.

(ii) A intuitional condition $P_a > P_b$ can be obtained when the attacks are not conducted on the training dataset, i.e, $\delta = \oslash$.

(iii) As indicated in Theorem 1, when P_a rises, the optimal type II error also increases for backdoor triggers δ. Consequently, in the simplified scenario where $P_a + P_b = 1$ and the robustness condition is defined as $P_1(\Gamma(\mathcal{I}); \Phi_1) > 0.5$, the distribution shift induced by δ can intensify. Consequently, as the smoothed classifier gains more confidence, the region of robustness expands.

A.2 Lemma and Proof

As defined in Subsection 3.3, given the distributions of null Φ_0 and alternative Φ_1. For convenience, the probabilities for type I and type II error are simplified as $\mathcal{E}_0$ and $\mathcal{E}_1$ below. For a given test sample $\mathcal{X}$, let ρ be a deterministic function such that $\rho(\mathcal{X}) \in [0, 1]$. Hypothesis testing aims to assess whether $\mathcal{X}$ is derived from either Φ_0 or Φ_1 distributions via a likelihood ratio test (LRT), we define the test ratio $\Theta_{\mathcal{X}} = f_{\Phi_1}(\mathcal{X})/f_{\Phi_0}(\mathcal{X})$ and choose a constant j such that:

$$\rho(\mathcal{X}) = \begin{cases} 0 & \text{if } \Theta_{\mathcal{X}} < i \\ j & \text{if } \Theta_{\mathcal{X}} = i \\ 1 & \text{if } \Theta_{\mathcal{X}} > i \end{cases} \tag{13}$$

where ρ refers to the probability of making type I error, i.e., rejecting the null hypothesis against the alternative hypothesis. The parameters i and j results in the bound-level optimal significance level ε_0 stated in hypothesis testing in Subsection 3.3 in full paper that satisfy $\mathcal{E}_0(\rho, \mathcal{X}) = j \cdot \Phi_0(\Theta_{\mathcal{X}} = i) + \Phi_0(\Theta_{\mathcal{X}} > i) = \varepsilon_0$.

Given that $\epsilon_0 \sim \Phi_0$ and $\epsilon_1 \sim \Phi_1$ be two random variables with densities f_0 and f_1 respectively, the likelihood ratio is denoted by $\Theta_{\mathcal{X}} = f_{\Phi_0}(\mathcal{X})/f_{\Phi_1}(\mathcal{X})$. Let $i_k = \inf\{\mathcal{E}(\Theta(\epsilon_0) \leq i) \geq k\}$ where $k \in [0, 1]$ and $i \geq 0$. Then we can obtain:

$$\mathcal{E}(\Theta(\epsilon_0) < i_k) \leq k \leq \mathcal{E}(\Theta(\epsilon_0) \leq i_k) \tag{14}$$

We provide proof for from left to right. From the left side of the inequality, we aim to firstly prove $\mathcal{E}(\Theta(\epsilon_0) < i_k) \leq k$. Given a set $M_n = \mathcal{X} : \Theta(\mathcal{X}) < i_k - 1/n$, while $M = \mathcal{X} : \Theta(\mathcal{X}) < i_k$. Assume that $\mathcal{X} \in \cup_n M$, $\exists n$ such that $\Theta(\mathcal{X}) < i_k - 1/n < i_k$, thus $\mathcal{X} \in \mathcal{M}$. Conversely, if $\mathcal{X} \in \mathcal{M}$, let $n \to \infty$ then $\Theta(\mathcal{X}) < i_k - 1/n$. Thus $\mathcal{X} \in \cup_n \mathcal{M}_n$ and $\mathcal{M} = \mathcal{M}_n$. Following by the above, we can obtain that $\mathcal{E}(\epsilon_0 \in M_n) = \mathcal{E}(\Theta(\epsilon_0) < i_k - 1/n) < k$. Since $M_n \subseteq M_{n+1}$, we can conclude that $\mathcal{E}(\Theta(\epsilon_0) < i_k) = \lim_{n\to\infty} \mathcal{E}(\Theta(\epsilon_0 \in M_n) < k$.

Regarding the right side, this follows directly from the definition of i_k if we show that the function $i \mapsto \mathcal{E}(\Theta(\epsilon_0) \leq i)$ is right-continuous. Given that $i \geq 0$ and $\{i_1, i_2, \cdots, i_n\}$ is a sequence in $R_{\geq 0}$ such that $i_n \to i$ as i decreases. Define the sets $I_n = \{\mathcal{X} : \Theta(\mathcal{X}) \leq i_n\}$ and note that $\mathcal{E}(\Theta(\epsilon_0) \leq i_n) = \mathcal{E}(\epsilon_0 \in I_n)$. Assume that $\mathcal{X} \in \{\mathcal{X} : \Theta(\mathcal{X}) \leq i\}$, there exits $\forall n$ we have $\Theta(\mathcal{X}) \leq i \leq i_n$ and $\mathcal{X} \in \cap_n I$ is obtained. Conversely, if $\mathcal{X} \in \cap_n I$, there exits $\forall i$ we have $\Theta(\mathcal{X}) \leq i_n \to i$ when $n \to \infty$. Thus $\cap_n I = \{\mathcal{X} : \Theta(\mathcal{X} \leq i)\}$, then $\lim_{n\to\infty} \mathcal{E}(\Theta(\epsilon_0) \leq i_n) = \mathcal{E}(\epsilon_0 \leq i)$ which is deduced based on $\lim_{n\to\infty} \mathcal{E}(\epsilon_0 \in I_n) = \mathcal{E}(\epsilon_0 \in \cap_n I)$. Therefore, $i \to \mathcal{E}((\epsilon_0 \leq i$ is right continuous, such that $i \leq \mathcal{E}(\Theta(\epsilon_0) \leq i_k)$.

Given that $\epsilon_0 \sim \Phi_0$ and $\epsilon_1 \sim \Phi_1$ be two random variables with densities f_0 and f_1 with relative to parameter u respectively, the likelihood ratio Θ is denoted by $\Theta_{\mathcal{X}} = f_{\Phi_0}(\mathcal{X})/f_{\Phi_1}(\mathcal{X})$ and ρ is a function mapping ensures $\rho(\mathcal{X}) \in [0,1]$ defined in 13. Let the probability of making type I and type II errors be $\mathcal{E}_0$ and $\mathcal{E}_1$ respectively, and $\rho^{'}$ denotes any other test. Then we have:

$$\mathcal{E}_0(\rho^{'}) \geq 1 - \mathcal{E}_0(\rho) \implies 1 - \mathcal{E}_1(\rho) \geq \mathcal{E}_1(\rho^{'}) \tag{15}$$

We use backward inference to prove $1-\mathcal{E}_1(\rho)-\mathcal{E}_1(\rho^{'}) \geq 0$ as below, given that $\mathcal{E}_0(\rho^{'})-(1-\mathcal{E}_0(\rho)) \geq 0$:

$$\begin{aligned}
1 - \mathcal{E}_1(\rho) - \mathcal{E}_1(\rho^{'}) &= \int_{\Theta>i} \rho^{'} f_1 du + \int_{\Theta\leq i} (\rho^{'} - 1) f_1 du + j \int_{\Theta=i} f_1 du && (16)\\
&= \int_{\Theta>i} \rho^{'} \Theta f_0 du + \int_{\Theta\leq i} (\rho - 1) \Theta f_0 du + j \int_{\Theta=i} \Theta f_0 du && (17)\\
&\geq i \cdot [\int_{\Theta>i} \rho^{'} \Theta f_0 du + \int_{\Theta\leq i} (\rho - 1) \Theta f_0 du + j \int_{\Theta=i} \Theta f_0 du] && (18)\\
&= i \cdot [\mathcal{E}_0(\rho^{'}) - (1 - \mathcal{E}_0(\rho))] \geq 0 && (19)
\end{aligned}$$

Given that ϵ_0 follows the distribution Φ_0 and ϵ_1 follows the distribution Φ_1, both of which are random variables with probability densities f_0 and f_1 respectively, relative to parameter u, the likelihood ratio Θ is defined as $\Theta_{\mathcal{X}} = f_{\Phi_0}(\mathcal{X})/f_{\Phi_1}(\mathcal{X})$. Additionally, ρ is a function that maps $\mathcal{X}$ to a value in the interval $[0,1]$, as defined in Equation 13. Let $\mathcal{E}_0$ and $\mathcal{E}_1$ represent the probabilities of making type I and type II errors respectively, and let ρ' denote any other test. Then we have:

$$\mathcal{E}_0(\rho^{'}) \leq \mathcal{E}_0(\rho) \implies \mathcal{E}_1(\rho) \geq \mathcal{E}_1(\rho^{'}) \tag{20}$$

Similar with Proof A.2, we also use backward inference to prove $\mathcal{E}_1(\rho) - \mathcal{E}_1(\rho^{'}) \geq 0$ as follows, given that $\mathcal{E}_0(\rho^{'}) - \mathcal{E}_0(\rho) \leq 0$:

$$\begin{aligned}
\mathcal{E}_1(\rho) - \mathcal{E}_1(\rho^{'}) &= \int_{\Theta>i} \rho^{'} f_1 du + \int_{\Theta\leq i} (\rho^{'} + 1) f_1 du - j \int_{\Theta=i} f_1 du && (21)\\
&= \int_{\Theta>i} \rho^{'} \Theta f_0 du + \int_{\Theta\leq i} (\rho^{'} + 1) \Theta f_0 du - j \int_{\Theta=i} \Theta f_0 du && (22)\\
&\geq i \cdot [\int_{\Theta>i} \rho^{'} \Theta f_0 du + \int_{\Theta\leq i} (\rho^{'} + 1) \Theta f_0 du - j \int_{\Theta=i} \Theta f_0 du] && (23)\\
&= i \cdot [\mathcal{E}_0(\rho) - \mathcal{E}_0(\rho^{'})] \geq 0 && (24)
\end{aligned}$$

A.3 Proof of Theorem 3.1

Let randomized smoothing follows the distribution $\eta : (\varepsilon, \epsilon)$ and the distribution of backdoored dataset with randomized noise is denoted as $\eta^{'} : (\varepsilon, \epsilon) + (\delta_{X^*}, \delta)$. Then based on 13, given a test sample $\mathcal{X}$, the likelihood ratio testing is $\Theta = f_{\eta^{'}}(\mathcal{X})/f_\eta(\mathcal{X})$. Consider the likelihood test ρ_a the corresponding confidence level is $1 - P_a$. Then we have:

$$i_k = \inf\{i \geq 0 : \mathcal{E}(\Theta(\eta) \leq i) \geq k)\} \tag{25}$$

and

$$j_k = \begin{cases} \frac{\mathcal{E}(\Theta(\eta) \leq i_k) - k}{\mathcal{E}(\Theta(\eta) = i_k)} & \text{if } \mathcal{E}(\Theta(\eta) = i_k \neq 0 \\ 0 & \text{if } \mathcal{E}(\Theta(\eta) = i_k = 0 \end{cases} \tag{26}$$

Originated from Lemma A.2, we can further obtain that $\mathcal{E}(\Theta(\eta) \leq i_k) = \mathcal{E}(\Theta(\eta) < i_k) + \mathcal{E}(\Theta(\eta) = i_k) \leq k + \mathcal{E}(\Theta(\eta) = i_k)$. Consider parameters for ρ_k as $i = i_k$ and $j = j_k$. Here the probability of ρ_k for making type I error is $\mathcal{E}(\rho_k) = 1 - k$, where $k \in [0,1]$. Similarly, if the likelihood test $\rho_a = \rho_{P_a}$, then $\mathcal{E}(\rho_a) = 1 - P_a$. Therefore, regarding equation 9 in full paper, $\mathcal{E}(f(X + \varepsilon, D^* + \epsilon) = Y_a) = f^*(Y_a \mid$

$X, D^*) \geq 1 - \rho_a$, combining the lemma A.2, $\rho(\mathcal{X}) = I_{f(X+\varepsilon, D+\epsilon)=Y_a}(\mathcal{X})$ and the likelihood ratio test for testing the null against the alternative $\rho = \rho_a$, we have:

$$f^*(Y_a \mid X + \delta_X, D + \delta) = 1 - \mathcal{E}_1(\rho) \geq \mathcal{E}_1(\rho_a) \tag{27}$$

Likewise, the likelihood test $\rho_b = \rho_{1-P_b}$, then $\mathcal{E}(\rho_b) = P_b$. Therefore, regarding equation 9 in full paper, for an arbitrary $Y \neq Y_a$, we can obtain $\mathcal{E}(f(X + \varepsilon, D^* + \epsilon) = Y) = f^*(Y \mid X, D^*) \leq P_b = \mathcal{E}_0(\rho_{b)}$. Following this, combining the lemma A.2, $\rho(\mathcal{X}) = I_{f(X+\varepsilon, D+\epsilon)=Y}(\mathcal{X})$ and the likelihood ratio test for testing the null against the alternative $\rho = \rho_b$, we have:

$$f^*(Y \mid X + \delta_X, D + \delta) = 1 - \mathcal{E}_1(\rho) \leq 1 - \mathcal{E}_1(\rho_b) \tag{28}$$

Therefore, $\mathcal{E}_1(\rho_a) - \mathcal{E}_1(\rho_b) > 1$, we can conclude that $f^*(Y_a \mid X + \delta_X, D + \delta) > \max_{Y \neq Y_a} f^*(Y \mid X + \delta_X, D + \delta)$.Thus it is guaranteed that $Y_a = \arg\max_Y f^*(Y \mid \mathcal{X} + \delta_{\mathcal{X}}, D^* + \delta)$.

A.4 Proof of Theorem 3.2

Stemming from the optimality of the likelihood ratio test, we show in Theorem 2 that this condition is tight. We show tightness by constructing a base classifier f, such that the smoothed classifier f^* is consistent with the class probabilities in equation 9 in full paper. For a given (fixed) input $(\mathcal{X}, \mathcal{D})$ but whose smoothed version is not robust for adversarial perturbations δ that violate in equation 10 in full paper. Let ρ_a and ρ_b be two likelihood ratio tests for testing the null $(\varepsilon, \epsilon) \sim \mathcal{E}_0$ against the alternative $(\varepsilon, \epsilon) + (\delta_X, \delta) \sim \mathcal{E}_1$ and let ρ_a be such that $\alpha(\rho_a) = 1 - P_a$ and ρ_b such that $\alpha(\rho_b) = P_b$. Since δ violates equation 10, we have that $\beta(\rho_a) + \beta(\rho_b) \leq 1$. Let p^* be given by:

$$p^*(Y \mid \mathcal{X}, \mathcal{D}) = \begin{cases} 1 - \rho_a(\mathcal{X} - X, \mathcal{D} - \mathcal{D}_*) & \text{if } Y = Y_a \\ \rho_b(\mathcal{X} - X, \mathcal{D} - \mathcal{D}_*) & \text{if } Y = Y_b \\ \frac{1 - p^*(Y_a \mid \mathcal{X}, \mathcal{D}) - p^*(Y_b \mid \mathcal{X}, \mathcal{D})}{|\mathcal{Y}| - 2} & \text{otherwise.} \end{cases}$$

In the given context, $\mathcal{D} - \mathcal{D}*$ signifies the subtraction operation performed on the features but not on the labels. It is noteworthy that for binary classification, where $|\mathcal{Y}| = 2$, it follows that $\rho a = \rho_b$, thus making p well-defined. This is because, in this scenario, the assumption is $P_a + P_b = 1$. However, if $|\mathcal{Y}| > 2$, it's immediate from the definition of p that $\sum_k p(Y \mid X, \mathcal{D}) = 1$. It's worth noting, from the derivation of ρ_a and ρ_b in the proof of Theorem 1, that (pointwise) $\rho_a \geq \rho_b$ holds provided $P_a + P_b \leq 1$. Consequently, for $Y \neq Y_a, Y_b$, it follows that $P(Y \mid x, \mathcal{D}) \propto \rho_a - \rho_b \geq 0$. Hence, p emerges as a well-defined conditional probability distribution over labels, and $f(X, \mathcal{D}) = \arg\max_Y p(Y \mid X, \mathcal{D})$ serves as a base classifier.

Moreover, to illustrate the consistency of the corresponding smoothed classifier f with the class probabilities stated in equation 9, consider:

$$f(Y_a \mid \mathcal{X}, \mathcal{D}) = \mathcal{E}(1 - \rho_a(X, D)) = P_a \; f(Y_b \mid X_0, \mathcal{D}0) = \mathcal{E}(\rho b(X, D)) = P_b \tag{29}$$

Furthermore, for any $Y \neq Y_a, Y_b$, we have $f(Y \mid \mathcal{X}, \mathcal{D}) = (1 - P_a - P_b)/(\mid Y \mid -2) \leq P_b$, since the assumption is $P_a + P_b \geq 1 - (\mid \mathcal{Y} \mid -2) \cdot P_b$. Hence, f aligns with the class probabilities in equation 9. Additionally, note that $f(Y_a \mid \mathcal{X} + \delta_{\mathcal{X}}, \mathcal{D} + \delta) = 1 - \beta(\rho_a) < \beta(\rho_b)$, demonstrating that indeed $Y_a \neq f^*(X_0 + \delta_{\mathcal{X}}, \mathcal{D} + \delta)$.