Appendix

This is the appendix for the submission: *Certified Robustness to Word Substitution Attack with Differential Privacy*. Section A contains additional proofs that are omitted in the paper; Section B presents additional experiment results; Section C provides additional details of the experiment.

A Proof

A.1 Proof for Lemma 4.1

Proof. Take y_1 as an example and $y_1 \in [0, 1]$.

$$\begin{aligned}
\mathbb{E}[f_{\mathcal{A}}^{y_1}(\mathbf{X})] &= \int_0^1 \mathbb{P}(f_{\mathcal{A}}^{y_1}(\mathbf{X}) > t)dt \\
&\overset{(a)}{\leq} e^{\epsilon}(\int_0^1 f_{\mathcal{A}}^{y_1}(\mathbf{X}') > t)dt) \\
&= e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_1}(\mathbf{X}')],
\end{aligned} \tag{15}$$

where (a) is the by definition of DP. The same proof holds for any $y \in \mathcal{Y}$. □

A.2 Proof for Lemma 4.2

Proof. By eq.(3) in the paper, $\forall \mathbf{X}' \in \mathscr{S}(L)$ we have:

$$\mathbb{E}[f_{\mathcal{A}}^{y_n}(\mathbf{X})] \leq e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] \tag{16}$$

$$\mathbb{E}[f_{\mathcal{A}}^{y_i}(\mathbf{X}')] \leq e^{\epsilon}\mathbb{E}[f_{\mathcal{A}}^{y_i}(\mathbf{X})], \quad i \neq n \tag{17}$$

Then we have:

$$\begin{aligned}
\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] &\overset{Eq(16)}{\geq} \frac{\mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X})]}{e^{\epsilon}} \\
&\overset{eq(4)}{\geq} \frac{e^{2\epsilon}\max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}))}{e^{\epsilon}} \\
&= e^{\epsilon}\max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X})) \\
&\overset{eq(17)}{\geq} \max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}'))
\end{aligned} \tag{18}$$

$$\forall \mathbf{X}' \in \mathscr{S}(\mathbf{X}, L), \mathbb{E}[f_{\mathcal{A}}^{y_c}(\mathbf{X}')] \geq \max_{i:i\neq c}\mathbb{E}(f_{\mathcal{A}}^{y_i}(\mathbf{X}')). \tag{19}$$

the definition of robustness at $\mathbf{X}$ holds. □

B Addition Experimental Results

B.1 Parameter Impact: Sampling Rate

As discussed in Section 4.3, we use Monte Carlo sampling to estimate the expected value of the randomized scoring function $\widehat{\mathbb{E}}[\mathcal{A}(\mathbf{X}')]$. The draw times n can influence the final estimation of the expected value. Here we present the result of tuning n on IMDB dataset. As shown in Table 2, we evaluate different n on two settings, $L = 3, \epsilon = 0.4$ and $L = 9, \epsilon = 1$. CP and CA represent certified percentage and certified accuracy respectively. With the increase of draw times n, while the certified percentage does not change significantly, the certified accuracy increases. The best certified accuracy is achieved when $n = 1000$.

C Experiment Details

C.1 Datasets

The detailed comparison between IMDB and AGNews are shown in Table 3.

	L = 9, eps = 1		l = 3, eps = 1.2	
N	CP	CA	CP	CA
10	0.551	0.836	0.511	0.761
50	0.55	0.841	0.45	0.777
100	0.547	0.845	0.484	0.781
500	0.553	0.846	0.468	0.782
1000	0.549	**0.857**	0.479	**0.814**

Table 2: Certified percentage and certified accuracy under different n

Dataset	IMDB	AGNews
Training size	20,000	120,000
Testing size	2000	2000
Task	binary	four-class
Vocab size	116,839	114,096
Average synonym size	3.52	3.79
Sentence length	269.97±200.88	44.97±12.55
embedding_dims	100	100

Table 3: Summary of datasets

C.2 Target Model

The architecture we use for both datasets are single-layer LSTM model with hidden size of 128. The batch size for IMDB is 32 and for AGNews is 63. The word embedding dimension is 100 for both of the datasets. The loss functions are binary crossentropy loss for IMDB and categorical crossentropy for AGNews. For both datasets. Both dataset are trained with 30 epoches. The optimizer is Adam with learning rate 1×10^{-2}. Dropout rate is 0.3.

C.3 Attack Algorithm

The attack algorithm we use to generate adversarial examples is PWWS(Ren et al., 2019), which calculates the word replacement order based on both the word saliency and the classification probability, and uses WordNet to build synonym set and replace named entities (NEs) with similar NEs to flip the prediction. In the experiments, we randomly generate 2000 adversarial examples.