Private Language Models via Truncated Laplacian Mechanism

Tianhao Huang[1,2,3]*, Tao Yang[1,2,3]*, Ivan Habernal[4], Lijie Hu[1,2], Di Wang[1,2]†,
[1]Provable Responsible AI and Data Analytics (PRADA) Lab,
[2]King Abdullah University of Science and Technology, [3]Nankai University,
[4]Research Center Trustworthy Data Science and Security of the University Alliance Ruhr,
Faculty of Computer Science, Ruhr University Bochum,
{tianhao.huang, tao.yang, lijie.hu, di.wang}@kaust.edu.sa
ivan.habernal@ruhr-uni-bochum.de

Abstract

Deep learning models for NLP tasks are prone to variants of privacy attacks. To prevent privacy leakage, researchers have investigated word-level perturbations, relying on the formal guarantees of differential privacy (DP) in the embedding space. However, many existing approaches either achieve unsatisfactory performance in the high privacy regime when using the Laplacian or Gaussian mechanism, or resort to weaker relaxations of DP that are inferior to the canonical DP in terms of privacy strength. This raises the question of whether a new method for private word embedding can be designed to overcome these limitations.

In this paper, we propose a novel private embedding method called the high dimensional truncated Laplacian mechanism. Specifically, we introduce a non-trivial extension of the truncated Laplacian mechanism, which was previously only investigated in one-dimensional space cases. Theoretically, we show that our method has a lower variance compared to the previous private word embedding methods. To further validate its effectiveness, we conduct comprehensive experiments on private embedding and downstream tasks using three datasets. Remarkably, even in the high privacy regime, our approach only incurs a slight decrease in utility compared to the non-private scenario.

1 Introduction

The recent developments of deep learning have led to significant success in various tasks in Natural Language Processing (NLP), from next word prediction in mobile keyboards (Ramaswamy et al., 2019), to critical tasks like predicting patient health conditions from clinical records (Yao et al., 2019). However, such applications may always involve user-generated textual data as the training dataset, which contains sensitive information. To address privacy concerns, text anonymization (Anandan et al., 2012; Pilán et al., 2022) has been commonly used, which involves identifying sensitive attributes and replacing them with alternative values. Nevertheless, such heuristic approaches become ineffective as deep neural networks often tend to memorize training data, making them susceptible to information leakage about the training data (Shokri et al., 2017; Carlini et al., 2021, 2019). One way that takes into account the limitations of existing approaches is designing Differentially Private (DP) algorithms. Differential privacy (Dwork et al., 2006a) is resilient to arbitrary side information that might be available to attackers and has become a de facto method for private data analysis (Xiang et al., 2024; Wang et al., 2020; Xiang et al., 2023; Xue et al., 2021; Huai et al., 2019; Wang et al., 2023; Wang and Xu, 2020; Huai et al., 2020; Hu et al., 2022).

Recently, there has been significant research focusing on differentially private (DP) versions of word embedding from various perspectives (Yue et al., 2021; Feyisetan et al., 2019; Krishna et al., 2021; Feyisetan et al., 2020a; Xu et al., 2021a,c; Carvalho et al., 2021b,a; Habernal, 2021, 2022). However, there are still some shortcomings in these approaches. On the one hand, several works consider adding Laplacian or Gaussian noise to the embedding space to ensure DP (Habernal, 2021; Krishna et al., 2021; Habernal, 2022). However, these mechanisms suffer from high noise levels, resulting in low utility, especially in the high privacy regime when the privacy parameter (ϵ) is small. Moreover, these mechanisms can even alter the semantics of sentences (see Fig.1). On the other hand, there is a growing body of work that focuses on a relaxation of the canonical definition of DP, known as metric DP, which can achieve better performance. However, as a relaxed notion of DP, Metric DP cannot provide the same level of strong privacy guarantees as the canonical DP (Mattern

*Equal contribution. Part of the work was done as a research intern at PRADA Lab.
†Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3980–3993
November 12-16, 2024 ©2024 Association for Computational Linguistics

Comparison of Private Embedding

Original:	Oh and we came on a Saturday night around	11:30	for	context.	(→Privacy Leakage)
Trlaplace:	Oh and we came on a Saturday night around	9:30pm	for	<unk>	(→Private and Fluent)
Laplace:	Oh and we came on a Saturday night around	around	for	<unk>	(→Semantic Problem)
Gaussian:	Oh and we came on a Saturday night around	11:30	for	<unk>	(→Privacy Leakage)

Figure 1: An example of (private) text re-write for different mechanisms with $\epsilon = 0.1$.

et al., 2022a). This raises the question of *whether we can develop improved private word embedding mechanisms within the framework of canonical DP that can have comparable performance with existing metric DP-based methods*.

In this paper, we provide an affirmative answer to the previous question by proposing a novel private mechanism for word embedding. Our approach is inspired by the superior performance of the truncated Laplacian mechanism in one-dimensional space (Geng et al., 2020). However, it remains unclear whether this superiority can extend to high dimensional cases, as directly extending the one-dimensional truncated Laplacian mechanism is challenging. To bridge this gap, we develop a high dimensional truncated Laplacian mechanism (TrLaplace), which is a non-trivial extension of the one-dimensional case. Theoretically, we show that compared with Laplacian and Gaussian mechanisms for private word embedding, TrLaplace-based private embedding has a lower variance. Moreover, we also conduct intensive experiments on both private embedding and downstream tasks to show our approach significantly outperforms the previous DP-based methods in the high privacy regime, and it will not drop much accuracy and utility compared with the non-private case. Moreover, compared to the existing metric DP-based method, our mechanism has even better performance for privacy tests while also keeping comparable performance for downstream tasks.

2 Related Work

Recent years have seen substantial advancements in language models within differential privacy (DP) frameworks. Due to the space limit, here we only mention the existing literature on private word embedding. We refer the readers to the survey (Hu et al., 2024a) for more details.

Current research on private word embeddings can be broadly categorized into two approaches: original DP-based methods and metric DP-based methods. The seminal work in the original DP category by Lyu et al. (2020b) introduces a framework utilizing the Unary Encoding mechanism. This approach was subsequently refined by Plant et al. (2021). Further improvements were made by Lyu et al. (2020a), who proposed a dropout technique for perturbed embeddings to enhance downstream task fairness. However, Qu et al. (2021) identify a critical privacy issue in Lyu et al. (2020a), noting that it requires access to users' raw data for fine-tuning during the training phase. Other notable contributions include works by Krishna et al. (2021), Habernal (2021), and Alnasser et al. (2021), who explore privatizing word embeddings. Krishna et al. (2021) and Alnasser et al. (2021) propose ADePT, an auto-encoder-based DP algorithm. Unfortunately, Habernal (2021) points out that ADePT is not differentially private by thorough theoretical proof. Igamberdiev et al. (2022) address reproducibility by providing source code for DP Auto-Encoder methods. In this paper, we aim to improve the performance of the mechanisms in Igamberdiev et al. (2022).

In the realm of metric DP, Feyisetan et al. (2020b) first study this problem and provide a general perturbation-and-projection framework. Xu et al. (2020a) reconsider this problem setting, replacing the Euclidean distance with the Mahalanobis distance to improve the utility. Subsequently, Xu et al. (2021d) introduce the Vickrey mechanism to further refine the utility in the projection step. To address the limitations of the multivariate Laplace mechanism, Xu et al. (2021b) and Carvalho et al. (2021c) propose a Truncated Gumbel Noise method. Feyisetan and Kasiviswanathan (2021) tackle high-dimensionality issues using random projection. Additionally, Feyisetan et al. (2019) define hyperbolic embeddings and utilize the Metropolis-Hastings algorithm for sampling from hyperbolic distributions. More recently, Tang et al. (2020) explore differential privacy with varying privacy levels for different words. Arnold et al.

(2023a) introduce sense embeddings with a sense disambiguation step prior to noise injection, and Arnold et al. (2023b) address common semantic context issues in prior private embedding mechanisms. It is crucial to clarify that the objective of this work is the privatization of embedded outputs rather than the embedded methods themselves. The pre-trained initial embedding methods and the corresponding embeddings of all words in the vocabulary are treated as public knowledge. This distinction is significant because it allows us to perform projections without incurring additional privacy costs. By leveraging publicly available pre-trained embeddings, our method demonstrates how effective privacy-preserving techniques can be implemented while still utilizing existing resources.

3 Preliminaries

Differential Privacy is a data post-processing technique designed to ensure data privacy by adding confusion to potential attackers. Specifically, suppose there is one dataset noted as $\mathcal{D}$, and we change or delete one data record in this dataset which we call $\mathcal{D}'$. If the output distributions of $\mathcal{D}$ and $\mathcal{D}'$ are close enough, then we cannot distinguish these two distributions, i.e., we cannot infer whether the deleted or replaced data sample is really in this dataset. The formal details are given by (Dwork et al., 2006b). Note that in the definition of DP, adjacency is a key notion. One of the commonly used adjacency definitions is that two datasets S and S' are adjacent (denoted as $S \sim S'$) if S' can be obtained by modifying one record in S.

Definition 1 *Given a domain of dataset $\mathcal{X}$. A randomized algorithm $\mathcal{A} : \mathcal{X} \mapsto \mathcal{R}$ is (ε, δ)-differentially private (DP) if for all adjacent datasets S, S' with each sample is in $\mathcal{X}$ and for all $T \subseteq \mathcal{R}$, the following holds*

$$\Pr(\mathcal{A}(S) \in T) \leq \exp(\varepsilon) \Pr(\mathcal{A}(S') \in T) + \delta.$$

When $\delta = 0$, we call the algorithm $\mathcal{A}$ is ε-DP.

In this work, we adopt a similar setting to previous research on private word embedding (Feyisetan et al., 2020a; Xu et al., 2021a; Krishna et al., 2021). We consider a scenario where a user inputs a word w from a discrete fixed vocabulary $\mathcal{W}$. Our goal is to preserve the user's privacy with respect to her/his word. To achieve this goal, we aim to design an algorithm that accepts w as input and whose distribution of output is close to the case where $w' \in \mathcal{W}$ is the input, with $w' \neq w$ is any other word. From the attacker's perspective, based on the output, he cannot distinguish whether the user's input word is w or w' as their output distributions are almost the same. Formally, we have the following definition.

Definition 2 *Given a discrete vocabulary $\mathcal{W}$, a randomized algorithm $\mathcal{A} : \mathcal{W} \mapsto \mathcal{R}$ is word-level (ϵ, δ)-differentially private (DP) if for all pair of words $w, w' \in \mathcal{W}$ and for all $T \subseteq \mathcal{R}$ we have $\mathbb{P}(\mathcal{A}(w) \in T) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(w') \in T) + \delta$. When $\delta = 0$, we call the algorithm $\mathcal{A}$ is ϵ-DP.*

In this paper, we assume the user holds a sentence $s = w_1 w_2 \cdots w_n$ with n words. And we aim to design an (ϵ, δ)-DP algorithm, which is private w.r.t. each word w_i.

4 Private Embedding via Truncated Laplacian Mechanism

In this section, we will provide details of our method. Generally speaking, for each token w_i, to achieve DP, our approach consists of three steps. First, each token w_i is mapped to an d-dimensional pre-trained word embedding $\phi(w_i)$. And we perform a clipping step to get a clipped embedding:

$$\text{CLIPEmb}(w_i) = \phi(w_i) \min\{1, \frac{C}{\|\phi(w_i)\|_2}\}, \quad (1)$$

where the threshold $C > 0$ is a hyper-parameter. In the second step, we add some random noise to the clipped embedding vector to make it satisfies DP. Finally, we will perform the projection step by finding the nearest word $\hat{w}_i$ to the perturbed and clipped embedding vector within the embedding space:

$$\hat{w}_i = \arg\min_{w \in \mathcal{W}} \|\phi(w) - \text{CLIPEmb}(w_i) - \eta\|_2, \quad (2)$$

where η is the randomized noise we add in the second step. See Algorithm 1 for details.

It is notable that the goal of clipping is to make the ℓ_2-norm of embedding vector be bounded so that we can adding noise to ensure DP, such as the Laplacian mechanism or Gaussian mechanism (Dwork and Roth, 2014).

Theorem 1 (Laplacian Mechanism) *Suppose $\text{CLIPEmb}(\mathbf{w}) \in \mathbb{R}^d$ denote the clipped embedding vector with threshold C. Then the mechanism $\mathcal{A}_{lap}(w) = \text{CLIPEmb}(w) + \eta_1$ is ϵ-DP, where $\eta_1 = (\eta_{1,1}, \cdots, \eta_{1,d})$ and $\eta_{i,j}$ is drawn from a Laplacian Distribution Lap$(\frac{\Delta_1(f)}{\epsilon})$*

Algorithm 1 Privacy Preserving Mechanism

Input: String $s = w_1 w_2 \dots w_n$, clipping threshold C, privacy parameter $\epsilon > 0$.
Output: String $\hat{s} = \hat{w}_1 \hat{w}_2 \dots \hat{w}_n$.
1: **for all** $i \in \{1, \dots, n\}$ **do**
2: Sample η from the truncated Laplacian distribution in Theorem 3.
3: Obtain the perturbed clipped embedding $\mathbf{r}_i = \text{CLIPEmb}(w_i) + \eta$.
4: Let $\hat{w}_i = \text{Proj}\,(\mathbf{r_i})$ as in (2).
5: **end for**
6: **return** $\hat{s} = \hat{w}_1 \hat{w}_2 \dots \hat{w}_n$.

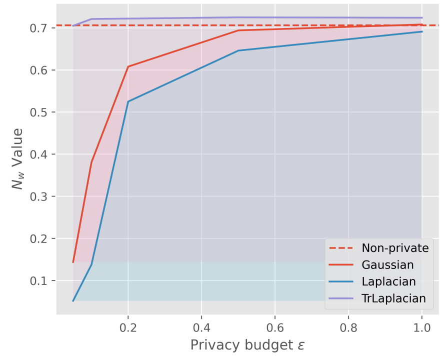


Figure 2: Privacy Test. Curves of the value N_w with privacy budget ϵ for Yelp dataset.

with $\Delta_1 = 2\sqrt{d}C$. *For a parameter* λ, *the Laplacian distribution has the density function* $Lap(\lambda)(x) = \frac{1}{2\lambda}\exp(-\frac{x}{\lambda})$.

Theorem 2 (Gaussian Mechanism) *Suppose* $\text{CLIPEmb}(\mathbf{w}) \in \mathbb{R}^d$ *denote the clipped embedding vector with threshold* C. *Then the mechanism* $\mathcal{A}_{gau}(w) = \text{CLIPEmb}(w) + \eta_2$ *is* (ϵ, δ)*-DP when* $\epsilon \leq 1$, *where* $\eta_2 \sim \mathcal{N}(0, \frac{8C^2 \ln(1.25/\delta)}{\epsilon^2} I_d)$ *is drawn from a Gaussian distribution.*

In the following we propose an improved mechanism namely high dimensional truncated Laplacian mechanism. Before that we first recall the probability density function of the one-dimensional truncated Laplacian distribution, which could be written as the following with some appropriate constants α, A and B:

$$f_{TLap}(x) = \begin{cases} \frac{1}{B} e^{-\alpha|x|}, & \text{for } x \in [-A, A] \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In our mechanism, we add high dimensional truncated Laplacian noise to the clipped embedding vector. Here each coordinate of the noise is i.i.d. sampled from a truncated Laplacian distribution with some specific α, A and B.

Remark 1 *It is notable that although using truncated Laplacian noise to ensure DP has been studied quite well (Geng et al., 2020; Sommer et al., 2021), all of them only considered the case where the dimension* $d = 1$ *and their methods cannot extend to the case where* $d > 1$. *For example, (Geng et al., 2020) only shows that adding noise with density function (3) with* $A = \frac{\Delta_1}{\epsilon}\log(1 + \frac{e^\epsilon}{2\delta})$ *and* $\alpha = \frac{\epsilon}{\Delta_1}$ *can ensure* (ϵ, δ)*-DP. Compared with the high dimensional case in Theorem 3 we can see the constant* A *is more complicated and the proof is also different. Thus, our mechanism cannot be considered as a trivial extension of the one-dimensional truncated Laplacian mechanism. Secondly, while the Laplacian mechanism can guarantee* ϵ*-DP, the truncated one can only ensure* (ϵ, δ)*-DP, which is the same as in the one-dimension case.*

However, as we will see below, our mechanism is superior to Laplacian mechanism for utility. It is also notable that we need to assume $\epsilon \leq 2\delta^{\frac{1}{d}}\sqrt{d}$, this is reasonable since we always wish ϵ to be as small as possible, as large ϵ indicates the algorithm is no longer private. If we want large $\epsilon > 2\delta^{\frac{1}{d}}\sqrt{d}$, we can use the trick of adding dummy dimension to the vector to increase its dimensionality manually and then projecting back to the original space after adding noise.

5 Theoretical Sensitivity Analysis

In the last section, we introduce our truncated laplacian mechanism, we will analyze its sensitivity and proof our claim in this section.

Theorem 3 *Suppose* $\text{CLIPEmb}(w) \in \mathbb{R}^d$ *is the clipped embedding vector with threshold* C. *Define* $\Delta_\infty = 2C$ *and* $\Delta_1 = 2\sqrt{d}C$. *For* $\epsilon \leq 2\delta^{\frac{1}{d}}\sqrt{d}$, *if*

$$\alpha = \frac{\epsilon}{\Delta_1}, A = -\frac{\Delta_1}{\epsilon}\log(1 - \frac{\epsilon}{2\delta^{\frac{1}{d}}\sqrt{d}})$$
$$B = \frac{2(1 - e^{-\alpha A})}{\alpha} = \frac{\Delta_\infty}{\delta^{\frac{1}{d}}},$$

then the mechanism $\mathcal{A}(w) = \text{CLIPEmb}(w) + \eta$ *is* (ϵ, δ)*-DP, where* $\eta = (\eta_1, \cdots, \eta_1)$ *and each* η_i *has the density function as in (3) with the above parameters.*

In the following, we will show our mechanism has lower variance than the Laplacian and Gaussian mechanism, which indicates that our method is superior theoretically.

Table 1: **Privacy Test.** Performance under fastText Embedding initialization for the non-private case ($\epsilon = \infty$) and three mechanisms (Gaussian, Laplacian and TrLaplacian) on Yelp dataset. The privacy budget ranges from 0.05 to 20. ↑ means a higher value under this metric indicates better results, and ↓ means the opposite. The best performance is **bolded**. The same symbols are used in the following tables by default.

	Original	Gaussian				Laplacian				TrLaplacian			
Privacy budget ϵ	∞	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5
Loss↓	3.35	35.01	29.33	9.31	4.50	36.23	29.69	17.15	5.58	**1.20**	**1.20**	**1.26**	**1.23**
Rouge1↑	87.8	12.72	28.68	77.95	86.90	10.99	27.96	58.97	85.16	**92.43**	**92.67**	**92.29**	**92.43**
BLEU↑	8.929	8.226	8.745	8.918	8.931	**8.998**	8.681	8.898	8.931	8.937	**8.938**	**8.937**	**8.938**
N_w ↑	0.713	0.138	0.232	0.661	0.765	0.058	0.225	0.484	0.753	**0.813**	**0.807**	**0.804**	**0.813**
BERT-S↑	0.967	0.864	0.873	0.945	0.966	0.857	0.867	0.908	0.962	**0.981**	**0.978**	**0.979**	**0.978**

	Original	Gaussian				Laplacian				TrLaplacian			
Privacy budget ϵ	∞	1	5	10	20	1	5	10	20	1	5	10	20
Loss↓	3.35	3.10	1.68	1.48	1.29	3.60	1.55	1.53	1.51	**1.22**	**1.25**	**1.28**	**1.27**
Rouge1↑	87.8	89.47	92.06	**92.40**	**92.49**	88.17	91.87	91.90	91.91	**92.42**	**92.35**	92.34	92.31
BLEU↑	8.929	8.936	8.937	8.936	8.936	8.935	8.937	8.936	8.934	**8.938**	**8.939**	**8.937**	**8.938**
N_w ↑	0.713	0.794	**0.809**	**0.804**	**0.813**	0.758	0.801	0.795	0.792	**0.807**	0.802	0.800	0.808
BERT-S↑	0.967	0.976	0.977	**0.978**	**0.980**	0.967	**0.978**	0.976	0.977	**0.979**	**0.978**	**0.978**	**0.980**

Theorem 4 *The variance of mechanism $\mathcal{A}$ in Theorem 3 is lower than the variance of Laplacian mechanism and Gaussian mechanism when $\delta \leq \frac{1}{e^d}$.*

6 Experiments

In this section, we conduct experiments for our method based on two parts: DP text re-write for fine-tuning (private embedding) and downstream tasks (sentiment analysis). We have open-sourced our code on `https://github.com/kaustpradalab/TrLap`.

6.1 Experimental Setup

Datasets. For the DP text re-write task, we use the Yelp * and Yahoo (Yang et al., 2019) datasets. The Yelp Open dataset is a subset of Yelp business, review, and user data with a training size of 8,539 and a testing size of 2,174. The Yahoo dataset contains 14,180 news articles and 34,022 click events. All data are collated to obtain a training, validation, and testing set segmented by sentences.

For downstream tasks, we use the SST-2 dataset (Socher et al., 2013) for the sentiment analysis task, from which we use 68,221 heavily polarized reviews from the Internet Movie Database. We divide the SST-2 dataset into an 80:20 ratio for training and testing. The training set consists of 54,576 reviews and the testing set consists of 13,645 reviews. We use the AG News dataset (Zhang et al., 2015) which includes news articles on the four main topics in the AG News corpus for the topic classification task. Following Meisenbacher et al. (2024), we randomly select a sample of 6,000 articles from each topic, totaling 16,000 for training, 4,000 for validation, and 4,000 for testing. The statistics of the datasets are presented in Tab. 4 in Appendix A.

Baseline. For DP text re-write, although Krishna et al. (2021) use the Laplacian mechanism to the sentence level DP instead of word level as in Definition 2. However, as Habernal (2021) mentions, the approach in Krishna et al. (2021) is not DP. Thus, here we will not compare with their method, and we will use the Laplacian and Gaussian mechanisms for the clipped embedding as baseline methods. For private fine-tuning, as we mentioned previously, all the previous methods only focus on metric DP instead of the original DP in Definition 2. Thus, our method is incomparable with theirs, and we will still use Laplacian and Gaussian mechanisms as baselines.

For utility test, we compare our method with Gaussian and Laplacian mechanism for the original DP notation, as well as Calibrated Multivariate Perturbations (CMP) (Feyisetan et al., 2020b), Mahalanobis Mechanism (Xu et al., 2020b) and Private Text Embedding (PTE) (Feyisetan and Kasiviswanathan, 2021), which are benchmark metric DP-based methods and have better utility than the original DP-based ones.

Evaluation Metrics. We use the loss of cross-entropy to measure the performance of language models. Specifically, cross-entropy is mainly used to determine how similar the actual output is to the expected output. Smaller model loss indicates less noise added to perturb the text. Addi-

*https://www.yelp.com/dataset/

Table 2: **Privacy Test.** Performance under GloVe Embedding initialization for the non-private case ($\epsilon = \infty$) and the three mechanisms, where the privacy budget ranges from 0.05 to 0.5.

		Original	**Gaussian**				**Laplacian**				**TrLaplacian**			
Privacy budget ϵ		∞	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5
Yahoo	**Loss↓**	2.95	51.25	26.66	9.92	5.97	51.43	37.86	15.35	7.31	**2.89**	**2.86**	**2.84**	**3.04**
	Rouge1↑	92.37	14.01	59.52	83.61	89.06	13.02	43.30	75.77	86.98	**92.44**	**92.43**	**92.41**	**92.25**
	BLEU↑	8.501	**9.286**	8.418	8.489	8.499	9.132	8.287	8.474	8.493	8.499	**8.500**	**8.497**	**8.504**
	N_w **↑**	0.703	0.072	0.511	0.595	0.628	0.066	0.334	0.566	0.642	**0.706**	**0.682**	**0.666**	**0.662**
	BERT-S↑	0.975	0.849	0.908	0.955	0.963	0.839	0.889	0.942	0.959	**0.976**	**0.971**	**0.971**	**0.971**
Yelp	**Loss↓**	3.07	34.67	21.62	10.61	5.98	36.00	34.64	14.86	7.38	**2.98**	**2.99**	**3.02**	**2.94**
	Rouge1↑	89.40	15.97	48.89	76.48	84.97	12.60	14.68	66.62	82.08	**89.45**	**89.47**	**89.34**	**89.54**
	BLEU↑	8.934	**8.976**	8.850	8.926	8.930	8.607	8.916	8.913	8.928	8.931	**8.935**	**8.936**	**8.936**
	N_w **↑**	0.706	0.144	0.381	0.608	0.694	0.052	0.138	0.525	0.646	**0.705**	**0.721**	**0.722**	**0.725**
	BERT-S↑	0.973	0.874	0.895	0.943	0.964	0.855	0.874	0.927	0.952	**0.971**	**0.973**	**0.971**	**0.972**

Table 3: **Privacy Test.** Performance under GloVe Embedding initialization for the non-private case ($\epsilon = \infty$) and the three mechanisms, where the privacy budget ranges from 1 to 20.

		Original	**Gaussian**				**Laplacian**				**TrLaplacian**			
Privacy budget ϵ		∞	1	5	10	20	1	5	10	20	1	5	10	20
Yahoo	**Loss↓**	2.95	4.28	3.01	3.03	2.98	4.93	3.24	3.05	3.13	**2.85**	**2.97**	**2.92**	**2.81**
	Rouge1↑	92.37	90.97	92.27	92.16	92.19	90.02	92.09	**92.28**	92.26	**92.41**	**92.35**	92.24	**92.45**
	BLEU↑	8.501	8.501	8.501	8.499	**8.500**	**8.503**	8.501	**8.502**	8.500	8.498	**8.501**	8.499	8.499
	N_w **↑**	0.703	0.637	0.680	0.664	0.672	0.660	0.658	0.675	0.655	**0.674**	**0.670**	**0.702**	**0.680**
	BERT-S↑	0.975	0.968	**0.973**	0.971	0.972	0.966	0.970	0.971	0.971	**0.974**	0.972	**0.975**	**0.974**
Yelp	**Loss↓**	3.07	4.74	3.14	3.13	2.97	5.02	3.30	3.66	3.17	**2.93**	**3.03**	**3.00**	**2.98**
	Rouge1↑	89.40	86.63	89.13	89.27	**89.80**	86.43	89.04	88.15	89.23	**89.68**	**89.40**	**89.37**	89.60
	BLEU↑	8.934	8.933	**8.936**	8.933	**8.944**	8.931	8.932	8.933	8.934	**8.934**	8.931	**8.934**	8.938
	N_w **↑**	0.706	0.708	**0.725**	0.708	0.739	0.691	0.721	0.704	0.699	**0.724**	0.700	**0.712**	**0.740**
	BERT-S↑	0.973	0.969	0.975	**0.975**	0.975	0.964	0.969	0.969	0.968	**0.975**	0.971	**0.976**	**0.976**

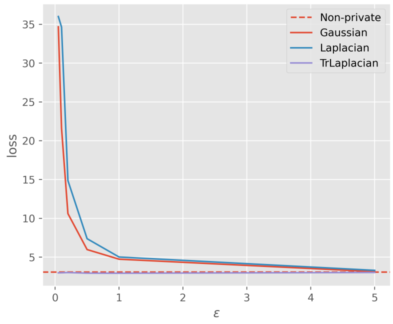
tionally, we will use Rouge1 and BLEU scores. Rouge1 (Lin, 2004) calculates recall using standard results and the number of 1-grams co-occurring in the auto-generated text. Similarly, BLEU (Papineni et al., 2002) measures the similarity between standard results and automatically generated text. Rouge1 measures word-level accuracy, while BLEU measures sentence fluency. Moreover, we use BERTScore (Zhang* et al., 2020) to measure the semantic similarity of the perturbed sentence with the original one. To measure the privacy-preserving ability, we use the percentage of N_w (Feyisetan et al., 2020a), which is the number of words that are not replaced. Thus, under the same privacy budget, larger N_w will be better (we want to change fewer words for accuracy).

Implementation Details. As an embedding can be considered as an initialization of the model, here we will consider three different initialization: Random embedding (Wieting and Kiela, 2019), GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). We conduct experiments on these embeddings and the subsequent fine-tuning in the DP model via our mechanism. Each pre-trained word embedding is a 300-dimensional vector, and the size of considered vocabulary is 10^4. For privacy budget, we set $\delta = \frac{1}{4^d}$, and we consider both the high privacy regime where $\epsilon \in \{0.05, 0.1, 0.2, 0.5\}$ and the low privacy regime $\epsilon \in \{1, 5, 10, 20\}$. For large ϵ we will use our previous dummy dimension trick ($d = 500$ for $\epsilon = 10$ and $d = 1700$ for $\epsilon = 20$).

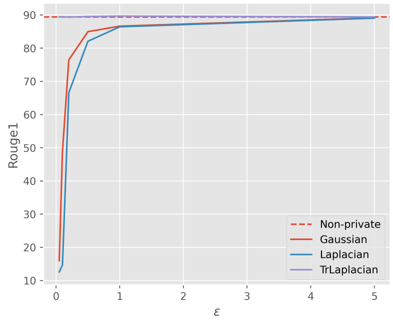
6.2 Privacy Experiment on Embedding

We first show the results on private embedding. Specifically, we use GloVe or fastText for the initialization, and then use three different private embedding mechanisms with different privacy budgets. Noted that large $\epsilon > 10$ is meaningless for privacy, we concentrate more on a small privacy budget in the main context.

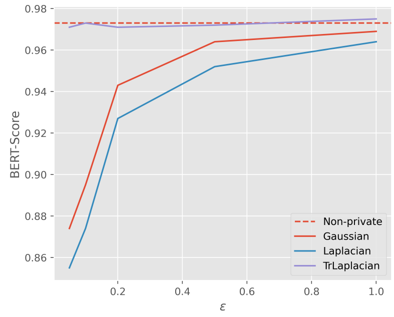
Fig. 1 and 6 show the text after projecting the clipped and perturbed embedding back to the word domain in step 4 of Algorithm 1 for different mechanisms when $\epsilon = 0.1$. We can see our method (TrLaplace) outperforms the other two methods from both privacy and semantic perspectives, while the Gaussian mechanism fails to obfuscate the time, and the Laplacian mechanism totally replaces the



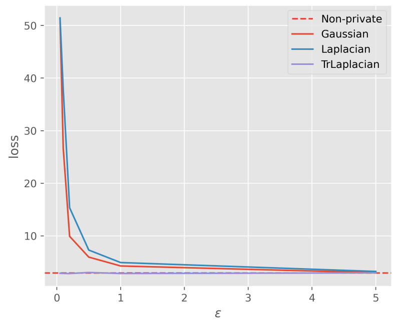
(a) Loss with $\epsilon = 0.05$ to $\epsilon = 5$.



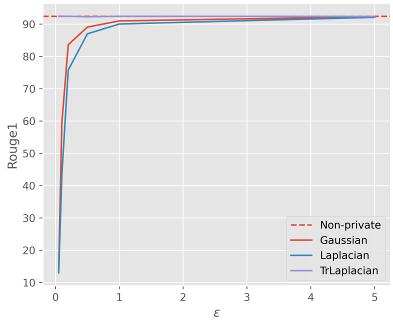
(b) Rouge1 with $\epsilon = 0.05$ to $\epsilon = 5$.



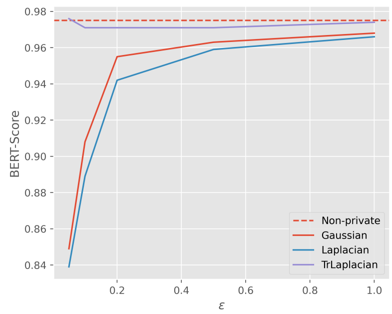
(c) BERT Score with $\epsilon = 0.05$ to $\epsilon = 1$.



(d) Loss with $\epsilon = 0.05$ to $\epsilon = 5$.



(e) Rouge1 with $\epsilon = 0.05$ to $\epsilon = 5$.



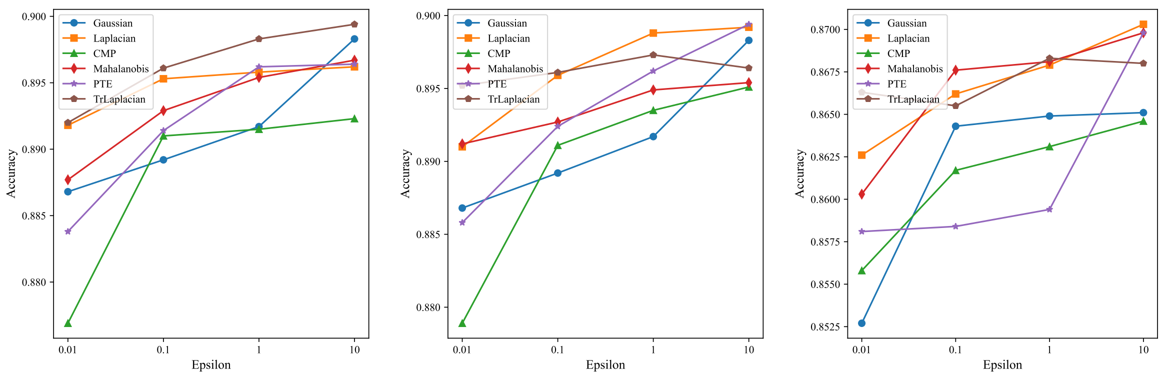
(f) BERT Score with $\epsilon = 0.05$ to $\epsilon = 1$.

Figure 3: **Privacy-Utility Test**. Curves of Loss, Rouge1 and BERTScore with different privacy budget ϵ for Yelp (Upper) and Yahoo (Lower) datasets.

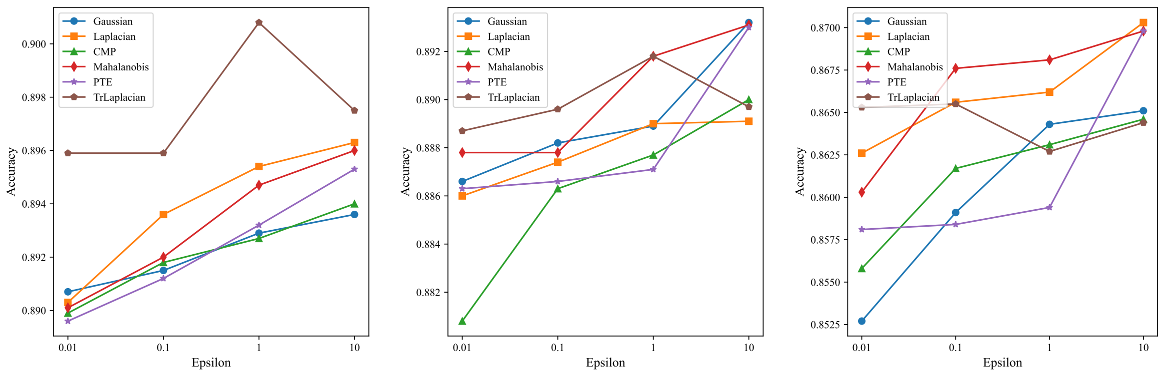
time by another word, which destroys the structure of the sentence.

Tab. 2 and Tab. 3 are the results on different metrics regarding private embedding with Glove initialization and Tab. 1 is with fastText initialization. We also present the detailed trends w.r.t ϵ for three mechanisms in Fig. 3. When $\epsilon < 1$, from Tab. 2 we can see that for both Yahoo and Yelp, the loss of Gaussian and Laplacian mechanisms will be catastrophically large while our mechanism has a much smaller loss. From Tab. 3 we can see we have almost the same phenomenon when in the low privacy regime. Moreover, for Rouge1, Trlaplacian also surpasses the other two mechanisms on both datasets, indicating that our mechanism consistently demonstrates superiority in lexical/syntactic aspects. For BLEU, the gap between all three mechanisms to the non-private case becomes small for both two datasets. But our method still has a slight advantage compared with the other two.

For N_w value, we can see in Fig. 2 and Fig. 5, our mechanism outperforms the other two mechanisms by changing less percentage of words to achieve the same privacy level, which indicates our method can exactly find sensitive words without hurting other words, thus keeps semantic properties. For BERTScore, our mechanism is almost the same as the non-private case, while there is a larger gap for others. It is notable that, in almost all experiments our mechanism is the best, and the Gaussian mechanism is better than the Laplacian mechanism, which matches our theorem. However, it becomes less obvious when ϵ is large. The main reason is that when ϵ is enough large the noise will be sufficiently small and becomes nearly negligible, which can also be supported by the proof of Theorem 4. For evaluation metrics, our mechanism may even be better than the non-private case, this may be due to small noise that could improve generalization, which is similar to adversarial training. Moreover. in real-world scenarios, the privacy budget must be very small as we just want to keep the word level of privacy. This is because, in more realistic scenarios, when we want to protect the whole sentence, the total privacy budget required to privatize an entire sequence may grow linearly with its length (due to the composition theorem). Thus, the budget for each word should be extremely small. This has also been mentioned in some previous work (Mattern et al., 2022b).



(a) SST-2, fastText (0.9013) (b) SST-2, GloVe (0.8997) (c) SST-2, Random (0.5607)



(d) AG News, fastText (0.9138) (e) AG News, GloVe (0.9161) (f) AG News, Random (0.2462)

Figure 4: Classification accuracy for all experimental settings. Each set of data is the average result of five experiments. We have included the baseline accuracy in parentheses in the subtitle of each subfigure.

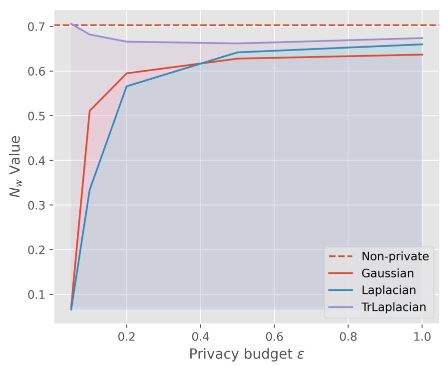


Figure 5: Privacy Test. Curves of N_w value w.r.t. privacy budget ϵ for Yahoo dataset.

6.3 Utility of Private Fine-tuning

We present the classification accuracy results for private fine-tuning across various embeddings and privacy levels in Fig. 4. While it is acknowledged that utility will be affected by the size of the budget, with a smaller budget potentially leading to lower utility, our experimental results show that our proposed method maintains relatively stable utility across different budget choices. Specifically, even in the high privacy regime ($\epsilon = 0.01$), our approach only incurs a slight decrease in utility compared to the non-private scenario. Similar capabilities of other methods can be observed in the experimental results of Meisenbacher et al. (2024).

We can also observe the effect of different pretrained embeddings. It is evident that when using GloVe and fastText pretrained embeddings, all methods achieve accuracy close to the baseline. However, when using random embeddings, the baseline accuracy is very low (0.5607 on SST-2, 0.2462 on AG News), but all methods significantly improve the accuracy of downstream sentiment analysis and topic classification tasks after training, with accuracy reaching up to about 0.87 when epsilon=10. This indicates that all methods are effective after training. Specifically, our method, Truncated Laplacian, maintains an accuracy of around 0.865 when trained on random embeddings, demonstrating that it maintains privacy while preserving the quality of embeddings, thereby offering excellent performance.

7 Conclusions

We introduce a novel method called the high dimensional truncated Laplacian mechanism for private embedding, which extends the one-dimensional case to the high-dimensional case. Theoretical analysis demonstrates that our method exhibits lower variance compared to existing private word embedding techniques. Experiments show that even in the high privacy regime, our approach incurs only a minimal loss in utility compared to the non-private case, which maintains privacy while preserving the quality of embeddings for promising performance.

8 Limitations

First, the word level DP has the disadvantages of length constraints and linear growth of privacy budget (Mattern et al., 2022a). However, such limitations are rooted in the definition of DP instead of our mechanism. Secondly, to ensure DP guarantees, in this paper, our mechanism involves clipping embedding vectors and adding calibrated noises, which inevitably introduce errors to the outputs of the task at hand. And these errors may affect different groups of individuals differently and may cause unfairness issues. However, we still need to mention that such unfairness issues are mainly due to the definition of DP rather than our method, as DP machine learning algorithms will always have a disparate impact on model accuracy (Bagdasaryan et al., 2019). Despite some limitations, word-level DP still offers unique advantages and potential applications (Hu et al., 2024b), and brings value to the DP-NLP community.

9 Ethics Review

This paper presents work whose goal is to advance the field of NLP. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

Di Wang and Lijie Hu are supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940. Ivan Habernal is supported by the Research Center Trustworthy Data Science and Security (`https://rc-trust.ai`), one of the Research Alliance centers within the UA-Ruhr (`https://uaruhr.de`).

References

Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. 2021. Privacy preserving text representation learning using BERT. In *Social, Cultural, and Behavioral Modeling - 14th International Conference, SBP-BRiMS 2021, Virtual Event, July 6-9, 2021, Proceedings*, volume 12720 of *Lecture Notes in Computer Science*, pages 91–100. Springer.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. t-plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023a. Driving context into text-to-text privatization. *CoRR*, abs/2306.01457.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023b. Guiding text-to-text privatization by syntax. *CoRR*, abs/2306.01471.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15453–15462.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pages 267–284. USENIX Association.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021a. BRR: preserving privacy of text data efficiently on device. *CoRR*, abs/2107.07923.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021b. TEM: high utility metric differential privacy on text. *CoRR*, abs/2107.07928.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021c. TEM: high utility metric differential privacy on text. *CoRR*, abs/2107.07928.

Rudrajit Das, Abolfazl Hashemi, Sujay Sanghavi, and Inderjit S. Dhillon. 2021. Dp-normfedavg: Normalizing client updates for privacy-preserving federated learning. *CoRR*, abs/2106.07094.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006a. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020a. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 178–186. ACM.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020b. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 178–186. ACM.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 210–219. IEEE.

Oluwaseyi Feyisetan and Shiva Kasiviswanathan. 2021. Private release of text embedding vectors. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 15–27.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. 2020. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 89–99. PMLR.

Ivan Habernal. 2021. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Habernal. 2022. How reparametrization trick broke differentially-private text representation learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 771–777. Association for Computational Linguistics.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024a. Differentially private natural language models: Recent advances and future directions. In *EACL (Findings)*, pages 478–499. Association for Computational Linguistics.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024b. Differentially private natural language models: Recent advances and future directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. 2022. High dimensional differentially private stochastic optimization with heavy-tailed data. In *PODS '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, pages 227–236. ACM.

Mengdi Huai, Di Wang, Chenglin Miao, Jinhui Xu, and Aidong Zhang. 2019. Privacy-aware synthesizing for crowdsourced data. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2542–2548. ijcai.org.

Mengdi Huai, Di Wang, Chenglin Miao, Jinhui Xu, and Aidong Zhang. 2020. Pairwise learning with differential privacy guarantees. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 694–701. AAAI Press.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-Rewrite: Towards Reproducibility and Transparency in Differentially Private Text Rewriting. In *The 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020a. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, Online. Association for Computational Linguistics.

Lingjuan Lyu, Yitong Li, Xuanli He, and Tong Xiao. 2020b. Towards differentially private text representations. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1813–1816. ACM.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022a. The limits of word level differential privacy. *CoRR*, abs/2205.02130.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022b. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024. A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off. *arXiv preprint arXiv:2404.03324*.

Radford M. Neal. 2003. Slice sampling. *The Annals of Statistics*, 31(3):705 – 767.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *CoRR*, abs/2202.00443.

Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. CAPE: Context-aware private embeddings for private language learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7970–7978, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving BERT. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 1488–1497. ACM.

Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. 2019. Federated learning for emoji prediction in a mobile keyboard. *CoRR*, abs/1906.04329.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Seattle, Washington, USA. Association for Computational Linguistics.

David M. Sommer, Lukas Abfalterer, Sheila Zingg, and Esfandiar Mohammadi. 2021. Learning numeric optimal differentially private truncated additive mechanisms. *CoRR*, abs/2107.12957.

Jingye Tang, Tianqing Zhu, Ping Xiong, Yu Wang, and Wei Ren. 2020. Privacy and utility trade-off for textual analysis via calibrated multivariate perturbations. In *Network and System Security - 14th International Conference, NSS 2020, Melbourne, VIC, Australia, November 25-27, 2020, Proceedings*, volume 12570 of *Lecture Notes in Computer Science*, pages 342–353. Springer.

Di Wang, Jiahao Ding, Lijie Hu, Zejun Xie, Miao Pan, and Jinhui Xu. 2023. Finite sample guarantees of differentially private expectation maximization algorithm. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 2435–2442. IOS Press.

Di Wang, Marco Gaboardi, Adam D. Smith, and Jinhui Xu. 2020. Empirical risk minimization in the non-interactive local model of differential privacy. *J. Mach. Learn. Res.*, 21:200:1–200:39.

Di Wang and Jinhui Xu. 2020. Principal component analysis in the local differential privacy model. *Theor. Comput. Sci.*, 809:296–312.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zihang Xiang, Tianhao Wang, Wanyu Lin, and Di Wang. 2023. Practical differentially private and byzantine-resilient federated learning. *Proc. ACM Manag. Data*, 1(2):119:1–119:26.

Zihang Xiang, Tianhao Wang, and Di Wang. 2024. Preserving node-level privacy in graph neural networks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4714–4732. IEEE.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021a. Density-aware differentially private textual perturbations using truncated gumbel noise. In *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021b. Density-aware differentially private textual perturbations using truncated gumbel noise. In *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020a. A differentially private text perturbation method using a regularized mahalanobis metric. *CoRR*, abs/2010.11947.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020b. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021c. On a utilitarian approach to privacy preserving text generation. *CoRR*, abs/2104.11838.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021d. On a utilitarian approach to privacy preserving text generation. *arXiv preprint arXiv:2104.11838*.

Zhiyu Xue, Shaoyang Yang, Mengdi Huai, and Di Wang. 2021. Differentially private pairwise learning revisited. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3242–3248. ijcai.org.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. 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 5077–5089, Hong Kong, China. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics Decis. Mak.*, 19-S(3):31–39.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A More Details and Experiments

Dataset. The statistics of dataset are shown in table 4.

Table 4: Dataset statistics used in this work.

Dataset	Avg. Length (tokens)	Train Size (neg/pos)	Test Size (neg/pos)
Yahoo	181	8539/8673	2174/2189
Yelp	19	3610/3310	909/912
SST-2	10	24214/30362	5994/7651
AG News	40	20000	4000

Table 5: **Time Cost.** Comparison of the time cost of each epoch (seconds) under GloVe Embedding initialization for the non-private case and three mechanisms (Gaussian, Laplacian and TrLaplacian), the privacy budget ranges from 0.05 to 20.

		$\epsilon < 1$				$\epsilon \geq 1$			
Privacy budget ϵ		0.05	0.1	0.2	0.5	1	5	10	20
Yahoo	**Non-private**	111							
	Gaussian	111	113	111	111	111	111	111	111
	Laplacian	111	113	111	111	111	111	111	111
	TrLaplacian	123	123	123	123	123	123	123	123
Yelp	**Non-private**	111							
	Gaussian	38	37	38	38	37	37	37	37
	Laplacian	38	37	37	37	37	37	37	37
	TrLaplacian	46	41	46	42	42	42	42	42

Implementation Details. Models in this paper are implemented based on the PyTorch [†] and TensorFlow [‡] with their libraries. Experiments are conducted on NVIDIA GeForce RTX 3090 GPUs. To

[†] https://pytorch.org/

[‡] https://www.tensorflow.org/

implement our mechanism, we use the acceptance-rejection sampling method (Neal, 2003) to sample a point from the high dimensional truncated Laplace distribution from the Laplace distribution, only by rejecting the samples outside the interval.

For text re-write, we use the auto-encoder model. The embedding is initialized with the 300-dimensional pre-trained Random, GloVe, and fastText word embedding. We use one-layer BiLSTM with dropout for the encoder, and using setup: dropout rate 0.5, Adam (Kingma and Ba, 2015) with an initial learning rate of 0.01 and betas (0.5, 0.999), batch size 1024, and number of training epochs 100. For the downstream classification task over the AG News and SST-2 dataset, we use Adam with an initial learning rate of 0.0001, a dropout rate of 0.2, and a batch size of 256. We set the maximum number of epochs to be 50.

B Omitted Proofs

Proof 1 (Proof of Theorem 3) *The proof is motivated by (Das et al., 2021). Consider a pair of tokens w, w'. Let perturbed encoder1 $r_1 = \mathrm{CLIPEmb}(w) + \eta_1$, also let $r_2 = \mathrm{CLIPEmb}(w') + \eta_2 = \mathrm{CLIPEmb}(w) + \boldsymbol{\Delta_s} + \eta_2$, where $\|\boldsymbol{\Delta_s}\|_1 \leq \Delta_1$ and $\|\boldsymbol{\Delta_s}\|_\infty \leq \Delta_\infty$ which are due to the clipping operation.*

Let us denote the set of possible values of r_k by $\mathcal{S}_k$ for $k = 1, 2$.

Define $\mathcal{U} = [-C - A, C + A]^d$. Note that for any subset $\mathcal{V} \subseteq \mathcal{U} - (\mathcal{S}_1 \cup \mathcal{S}_2)$, $\mathbb{P}(r_1 \in \mathcal{V}) = \mathbb{P}(r_2 \in \mathcal{V}) = 0$, hence (ϵ, δ)-DP is satisfied for this part. We need to ensure (ϵ, δ)-DP is satisfied for all elements in $\mathcal{S}_1 \cup \mathcal{S}_2$ too.

First, consider an element $s \in \mathcal{S}_1 \cap \mathcal{S}_2$. Then:

$$f(r_1 = s) = f(\eta_1 = s - \mathrm{CLIPEmb}(\mathbf{s}))$$

Similarly:

$$f(r_2 = s) = f(\eta_2 = s - \mathrm{CLIPEmb}(\mathbf{s}) - \boldsymbol{\Delta_s})$$

Using the above equations:

$$\begin{aligned}
&\exp(-\alpha\Delta_1) \leq \exp(-\alpha \|\Delta_s\|_1) \\
&\leq \frac{f(r_1 = s)}{f(r_2 = s)} \leq \exp(\alpha \|\Delta_s\|_1) \leq \exp(\alpha\Delta_1)
\end{aligned}$$

From the above equation, setting setting $\alpha = \epsilon/\Delta_1$ ensures pure ϵ-DP for all $s \in \mathcal{S}_1 \cap \mathcal{S}_2$. With this, it follows that for any $\mathcal{V} \subseteq \mathcal{S}_1 \cap \mathcal{S}_2$:

$$e^{-\epsilon}\mathbb{P}(r_2 \in \mathcal{V}) \leq \mathbb{P}(r_1 \in \mathcal{V}) \leq e^{\epsilon}\mathbb{P}(r_2 \in \mathcal{V}).$$

by setting $\alpha = \epsilon/\Delta_1$.

Now consider an element $s \in \mathcal{S}_2 - \mathcal{S}_1$. Clearly, $f(r_1 = s) = 0$. Also:

$$\max_{s \in \mathcal{S}_2 - \mathcal{S}_1} f(r_2 = s) \leq \frac{1}{B}.$$

But notice that volume$(\mathcal{S}_2 - \mathcal{S}_1) \leq \Delta_\infty^d$. This follows from the fact that for every coordinate, there are at most Δ_∞ levels that can be attained by r_2 but not by r_1. Thus, for any $\mathcal{T} \subseteq \mathcal{S}_2 - \mathcal{S}_1$, we have

$$\mathbb{P}(r_1 \in \mathcal{T}) = 0 \text{ and } \mathbb{P}(r_2 \in \mathcal{T}) \leq \left(\frac{\Delta_\infty}{B}\right)^d$$

Similarly, for any $\mathcal{T} \subseteq \mathcal{S}_1 - \mathcal{S}_2$, we have

$$\mathbb{P}(r_2 \in \mathcal{T}) = 0 \text{ and } \mathbb{P}(r_1 \in \mathcal{T}) \leq \left(\frac{\Delta_\infty}{B}\right)^d.$$

Now, let us now consider some general $\mathcal{T} \subseteq \mathcal{S}_1 \cup \mathcal{S}_2$. Let $\mathcal{T}_0 = \mathcal{T} \cap (\mathcal{S}_1 \cup \mathcal{S}_2)$, $\mathcal{T}_1 = \mathcal{T} \cap (\mathcal{S}_1 - \mathcal{S}_2)$ and $\mathcal{T}_2 = \mathcal{T} \cap (\mathcal{S}_2 - \mathcal{S}_1)$. It is easy to see that $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \cup \mathcal{T}_2$ and that $\mathcal{T}_0, \mathcal{T}_1$ and $\mathcal{T}_2$ are pairwise-disjoint. Then:

$$\begin{aligned}
\mathbb{P}(r_1 \in \mathcal{T}) &= \mathbb{P}(r_1 \in \mathcal{T}_0) + \mathbb{P}(r_1 \in \mathcal{T}_1) \\
&\quad + \mathbb{P}(r_1 \in \mathcal{T}_2) \\
&\leq e^{\epsilon}\mathbb{P}(r_2 \in \mathcal{T}_0) + \left(\frac{\Delta_\infty}{B}\right)^d + 0 \\
&\leq e^{\epsilon}\mathbb{P}(r_2 \in \mathcal{T}) + \left(\frac{\Delta_\infty}{B}\right)^d.
\end{aligned} \tag{4}$$

Thus, we can set $\delta = (\frac{\Delta_\infty}{B})^d$. Obviously, this result is only useful if $B > \Delta_\infty$.

For each coordinate

$$\begin{aligned}
&\int_{x \in \mathbb{R}} f_{\mathrm{TLap}}(x)dx = \int_0^A 2\frac{1}{B}e^{-\alpha|x|}dx \\
&= \frac{2}{B\alpha}\left(1 - e^{-\alpha A}\right) = 1
\end{aligned}$$

We can solve $B = \frac{2(1-e^{-\alpha A})}{\alpha}$. Thus, take $B = \frac{\Delta_\infty}{\delta^{\frac{1}{d}}}$, we can see $A = -\frac{1}{\alpha}\log(1 - \frac{\alpha\Delta_\infty}{2\delta^{\frac{1}{\delta}}}) = -\frac{\Delta_1}{\epsilon}\log(1 - \frac{\epsilon}{2\sqrt{d}\delta^{\frac{1}{\delta}}})$.

Proof 2 (Proof of Theorem 4) *We first show the variance of our mechanism $\mathcal{A}$ is bounded by $2\frac{d\Delta_1^2}{\epsilon^2}$.*

We can easily see that the variance is $\mathbb{E}\|\mathcal{A}(w) - w\|_2^2 = dV$ with $V = \int_{x \in \mathbb{R}} f_{\mathrm{TLap}}(x)|x|^2dx$, so

Comparison Semantic Problem of Private Embedding	
Original:	do not come here! food poisoning alert! (→Neg.)
Trlaplace:	do not come here! food poisoning alert! (→Neg.)
Laplace:	this place is awesome! love this place! (→Pos.)
Gaussian:	do not go here! food glorious <unk>! (→Pos.)

Figure 6: Another example of text re-write with different mechanisms with $\epsilon = 0.1$. The Gaussian and Laplacian mechanisms destroyed the semantic properties of the original sentence.

$$\begin{aligned}
&\int x^2 f(x)dx \\
=&2\frac{1}{B}\int_0^A e^{-\alpha x}x^2 dx \\
=&2\frac{1}{B}\int_0^A -\frac{1}{\alpha}x^2 d\left(e^{-\alpha x}\right) \\
=&2\frac{1}{B}(-\frac{1}{\alpha})A^2 e^{-\alpha A} + 2\frac{1}{B}\int_0^A \frac{1}{\alpha}e^{-\alpha x}2xdx
\end{aligned} \tag{5}$$

and

$$\begin{aligned}
&\int_0^A \frac{1}{\alpha}e^{-\alpha x}2xdx \\
=&-\int_0^A \frac{1}{\alpha^2}\cdot 2xd\left(e^{-\alpha x}\right) \\
=&-\frac{1}{\alpha^2}2Ae^{-\alpha A} + \int_0^A \frac{2}{\alpha^2}e^{-\alpha x}dx \\
=&-\frac{1}{\alpha^2}2A\cdot e^{-\alpha A} + \frac{2}{\alpha^3}\left(1-e^{-2\alpha A}\right)
\end{aligned} \tag{6}$$

Thus, we have

$$\begin{aligned}
V =& -2\frac{1}{\alpha}\frac{1}{B}A^2 e^{-\alpha A} - 4\frac{1}{\alpha^2}\frac{1}{B}Ae^{-\alpha A} \\
&+ 4\frac{1}{\alpha^3}\frac{1}{B}\left(1-e^{-\alpha A}\right) \\
&= -2\frac{1}{\alpha}\frac{1}{B}Ae^{-\alpha A}(A+2\frac{1}{\alpha}) + 2\frac{\Delta_1^2}{\varepsilon^2} \\
&< 2\frac{\Delta_1^2}{\varepsilon^2}.
\end{aligned} \tag{7}$$

Thus, in total we have $\mathbb{E}\|\mathcal{A}(w)-w\|_2^2 \leq \frac{2d\Delta_1^2}{\epsilon^2} = \frac{8d^2C^2}{\epsilon^2}$.

Next for Laplacian mechanism in Theorem 1 we have $\mathbb{E}[\|\mathcal{A}_{lap}(w)-w\|_2^2] = \frac{2d\Delta_1^2}{\epsilon^2}$. *Thus the variance of high dimensional truncated Laplacian is always lower than Laplacian.*

Similarly, the variance of Gaussian mechanism in Theorem 4 is $\frac{8C^2 d(\ln 1.25 + \ln 1/\delta)}{\epsilon^2}$, *we can easily see that our mechanism has lower variance when* $\delta \leq \frac{1}{e^d}$.