

# With Privacy, Size Matters: On the Importance of Dataset Size in Differentially Private Text Rewriting

Stephen Meisenbacher and Florian Matthes

Technical University of Munich

School of Computation, Information and Technology

Department of Computer Science

Garching, Germany

{stephen.meisenbacher, matthes}@tum.de

## Abstract

Recent work in Differential Privacy with Natural Language Processing (DP NLP) has proposed numerous promising techniques in the form of *text rewriting* mechanisms. In the evaluation of these mechanisms, an often-ignored aspect is that of *dataset size*, or rather, the effect of dataset size on a mechanism’s efficacy for utility and privacy preservation. In this work, we are the first to introduce this factor in the evaluation of DP text privatization, where we design utility and privacy tests on large-scale datasets with dynamic split sizes. We run these tests on datasets of varying size with up to one million texts, and we focus on quantifying the effect of increasing dataset size on the privacy-utility trade-off. Our findings reveal that dataset size plays an integral part in evaluating DP text rewriting mechanisms; additionally, these findings call for more rigorous evaluation procedures in DP NLP, as well as shed light on the future of DP NLP in practice and at scale.

## 1 Introduction

In the study of Differential Privacy in Natural Language Processing (DP NLP), a multitude of approaches have been proposed, ranging from word-level obfuscation mechanisms to document-level rewriting techniques (Klymenko et al., 2022; Hu et al., 2024). To evaluate the efficacy of a proposed DP rewriting mechanism, researchers often perform privacy evaluations, where a defined attack vector is shown to be mitigated through the privatization of textual data (Feyisetan et al., 2020; Mattern et al., 2022; Meisenbacher et al., 2024b). In addition, researchers may opt to demonstrate that a mechanism also preserves data utility, or usefulness in a downstream task (Weggenmann et al., 2022; Igamberdiev et al., 2022). These evaluations are based on a chosen set of datasets, which ideally simulate some plausibly “sensitive” scenario.

The primary factor in testing DP rewriting is the  $\epsilon$  parameter, also known as the *privacy budget* in

DP (Dwork, 2006). Evaluating a given mechanism over a range of  $\epsilon$  values is useful in investigating its behavior over various privacy regimes, and it is generally true that lower  $\epsilon$  values lead to higher privacy and lower utility, and vice versa (Abadi et al., 2016). A secondary factor that is highly important, yet often not considered, is that of *dataset size*, namely, how a mechanism behaves when run on datasets of varying size. While it is generally believed in DP that as dataset size grows, the effect of the injected random noise becomes less adverse (Ponomareva et al., 2023), there is little empirical evidence of this in the DP NLP literature. Above all, previous experiments do not consider differently sized splits of a dataset, therefore only presenting a static view of the performance of DP rewriting mechanisms.

We address this gap by investigating the impact of dataset size on the privacy and utility preservation capabilities of DP text rewriting methods, quantifying the shift in the privacy-utility trade-off as dataset size grows. We are guided by the following research question:

*What observations can be made regarding privacy and utility as dataset size grows in local DP text rewriting scenarios?*

To answer this, we design a comprehensive evaluation consisting of four large-scale datasets, with which utility and privacy experiments are conducted using five differently sized subsamples of each dataset. By doing so, we empirically demonstrate the observable effects of DP text rewriting as dataset size grows. We critically analyze these results, which give way to an outlook on the benefits and challenges of DP text privatization at scale, a point which is often not addressed in the literature.

Our findings reveal that dataset size does indeed matter with DP text rewriting, as mechanisms behave differently at various dataset splits and privacy budgets. In addition, we learn that as dataset size increases, trade-offs from privatization generally

become more favorable, making a case for DP text privatization at scale. Concretely, we make the following contributions to the field of DP NLP:

1. We are the first to conduct experiments in DP text rewriting in a *variable dataset size* setting.
2. We empirically demonstrate and highlight the importance of dataset size in DP rewriting.
3. We provide an outlook on future DP NLP work, focused on making it practical at scale.

## 2 Related Work

Work in DP NLP primarily investigates addressing the challenges of meaningful and effective text privatization while also finding a balance with preserved utility (Weggenmann et al., 2022). Several recent works propose novel DP mechanisms for NLP (Hu et al., 2024), often leveraging the *local* model of DP (Kasiviswanathan et al., 2008) for *text rewriting*, and others have focused on investigating the challenges of DP NLP (Feyisetan et al., 2021; Mattern et al., 2022; Klymenko et al., 2022; Meisenbacher et al., 2025), as well as highlighting important points for moving the field forward (Meisenbacher and Matthes, 2024; Vu et al., 2024). In particular, the evaluation and benchmarking of DP NLP is seen as crucial yet challenging (Igamberdiev et al., 2022; Meisenbacher et al., 2024b; Arnold, 2025), and recent works have proposed techniques for benchmarking DP text rewriting.

In previous works, evaluations often use publicly available datasets to mirror sensitive data scenarios. For example, datasets such as IMDb (Maas et al., 2011), Trustpilot (Hovy et al., 2015), or Yelp (Zhang et al., 2015), which contain some sensitive attributes (e.g., author ID or gender), are frequent choices for empirical privacy experiments. For utility evaluations, the choice of datasets is often varied (Meisenbacher et al., 2024b). In many recent local DP text rewriting works, the size of the evaluation datasets is quite small, particularly in privacy evaluations (Mattern et al., 2022; Meehan et al., 2022; Utpala et al., 2023). One notable exception is DP-BART (Igamberdiev and Habernal, 2023), which uses extensive datasets, yet without empirical privacy evaluations. Furthermore, in these works, the datasets are fixed in size, and the effect of privatization is not tested for different splits.

We view this to be a considerable gap, namely that DP text rewriting evaluations predominantly (1) do not generally test on larger-scale datasets,

thus leaving it unclear whether proposed methods are effective at scale, and (2) do not vary the size of a given evaluation dataset, thus leaving the impact of dataset size on DP text rewriting unexplored. As such, we address these two shortcomings in this work, with the goal of showing the merit of extended evaluation setups in future DP NLP works.

## 3 Experimental Setup and Results

To investigate the effect of dataset size in DP text rewriting, we design utility and privacy experiments where performance is measured on large datasets with various split sizes, described next.

### 3.1 Datasets and Tasks

In our initial experiments, we utilize four datasets with associated utility and privacy tasks.

**AG News.** The *AG News* dataset is a corpus prepared by Zhang et al. (2015) from a larger corpus of more than one million news articles. This subset contains 120k articles from four news domains, with 30k samples per class. The articles are shorter in nature, with an average word length of 43.95. We use this large-scale dataset to test the effect of DP text rewriting in a four-class classification task.

**MNLI.** The *Multi-Genre Natural Language Inference Corpus* (MNLI) is a dataset part of the GLUE benchmark (Wang et al., 2018), where MNLI contains *premise* and *hypothesis* pairs. The corresponding task is to classify whether the premise text *entails*, *contradicts*, or is *neutral* to the hypothesis text. We utilize the entire dataset of 392,702 samples, with an average premise length of 22.28 words. For DP rewriting, we only privatize the premise text, leaving the hypothesis intact.

**Trustpilot Reviews.** The *Trustpilot* corpus is a large-scale collection of user reviews. The corpus prepared by Hovy et al. (2015) tags each review with the stars provided (1-5), as well as the gender of the review writer. We utilize the *en-US* split, taking all reviews with gender information (male / female), and filtering by negative reviews (1-2 stars) and positive reviews (5 stars). This results in a dataset of 366,210 reviews with an average of 52.39 words per review. We use this dataset for a utility task (binary sentiment analysis), as well as an adversarial privacy task (gender inference).

**Yelp Reviews.** The *Yelp Open Dataset* is a massive corpus of nearly seven million user reviews

from Yelp. We take a smaller subset of the top-50 most frequently occurring users in the dataset<sup>1</sup>, which results in a dataset of 22,043 reviews with an average length of 196.29 words. This was done to facilitate a reasonable adversarial privacy task for authorship identification, which is paired with a utility task of binary sentiment analysis.

### 3.2 DP Rewriting Methods

For private rewriting of the chosen datasets, we choose three local DP rewriting mechanisms from the recent literature, which operate on three different levels of the syntactic hierarchy, i.e., yielding word-, token-, and document-level DP.

#### 1-DIFFRACTOR (Meisenbacher et al., 2024a).

The 1-DIFFRACTOR mechanism leverages metric local DP (MLDP) to obfuscate texts word-by-word in an efficient manner. We utilize the geometric version of the mechanism to rewrite texts with a word-level privacy budget of  $\varepsilon \in \{0.5, 1, 3\}$ , following values chosen in the original work.

**DP-PROMPT (Utpala et al., 2023).** The DP-PROMPT mechanism uses a temperature sampling mechanism during token sampling in text generation. We choose the FLAN-T5-LARGE model (Chung et al., 2022) as the underlying LM, and clip the logits to the range of  $(-95, 8)$ <sup>2</sup>, which are then normalized to the range of  $[0, 1]$ . Choosing temperature values of  $T \in \{1.75, 1.5, 1.25\}$ , thus resulting in *per-token* privacy budgets of  $\varepsilon \in \{1.14, 1.3, 1.6\}$ <sup>3</sup>, respectively.

#### DP-BART (Igamberdiev and Habernal, 2023).

The DP-BART mechanism rewrites texts on the *document-level* by adding DP noise to encoder representations in a BART model (Lewis et al., 2020). We use the base version of the mechanism (**DP-BART-CLV**) with a BART-BASE model and clipping range  $(-0.1, 0.1)$ . For document-level privatization, we choose the budgets of  $\varepsilon \in \{500, 1000, 1500\}$ , following the original work.

**A note on comparability.** We caution that the main focus of our experiments is to measure the effect of dataset size on privacy and utility metrics, and not to draw conclusions on the comparative effectiveness of the selected DP mechanisms. The latter would require a careful selection of privacy

budget parameters to ensure proper comparability between the three mechanisms operating on different linguistic levels and therefore offering differing privacy guarantees. Instead, we follow the  $\varepsilon$  values chosen by the original works, allowing for an analysis on the observable effects *within* a mechanism across dataset sizes and privacy budgets.

### 3.3 Experimental Procedure

First, each dataset is DP rewritten using the three chosen mechanisms and their privacy budgets, yielding nine private counterparts per dataset. These are then used to create five splits of various sizes: 10%, 25%, 50%, 75%, and 100%. This results in a total of 45 private “datasets” per original dataset, or 180 datasets in total.

Using these datasets, we first perform training and evaluation for the associated utility task for each dataset. To do this, we train a DEBERTA-V3-BASE (He et al., 2021) model on a 90% training split, and evaluate the trained model’s performance on the 10% val split. For utility tasks, we report the micro-F1 score from this evaluation.

The privacy experiments are conducted in a similar manner, where a DEBERTA-V3-BASE model is now trained as an adversarial classification model (2-class gender inference or 50-class author identification). Following previous work, we perform the privacy tests in both the *static* and *adaptive* settings (Weggenmann et al., 2022; Mattern et al., 2022; Utpala et al., 2023), where the static adversary trains on the original (non-privatized) texts and is evaluated on the private val split, while the adaptive adversary trains on the private train split.

In addition to reporting the micro-F1 of these evaluations, we also calculate the *relative gain* ( $\gamma$ ), which represents the trade-off between the paired utility and privacy tasks. For the calculation of  $\gamma$ , we define  $P_o, U_o$  to represent the baseline (non-rewritten) privacy and utility scores, respectively, and  $P_r, U_r$  be the scores observed on the privatized datasets. With this, relative gain is defined as  $\gamma = (U_r/U_o) - (P_r/P_o)$ , where higher scores are better. Note that we calculate the change in micro-F1 over majority-class guessing on the validation set for utility tasks, denoted  $MG_u$  (utility), as the Trustpilot and Yelp datasets sentiment analysis tasks are imbalanced (positive negative); thus  $RG = \frac{U_r - MG_u}{U_o - MG_u} - \frac{P_r}{P_o}$ . The exact procedure for calculating  $\gamma$  values is outlined in Appendix A.

We also design a *indistinguishability* test to measure the effect of dataset size on lending plausible

<sup>1</sup><https://business.yelp.com/data/resources/open-dataset/>

<sup>2</sup>Based on an empirical measurement of 100 randomly selected dataset texts, taking  $(logit\_mean, logit\_mean + 4 \cdot logit\_std)$ .

<sup>3</sup>Following  $\varepsilon = \frac{2\Delta}{T}$ , where the sensitivity ( $\Delta$ ) here is 1.

Split %	Baseline	AG News										MNLI										
		1-DIFFRACTOR			DP-PROMPT			DP-BART				1-DIFFRACTOR			DP-PROMPT			DP-BART				
0.5	1	3	1.14	1.3	1.6	500	1000	1500	Baseline	0.5	1	3	1.14	1.3	1.6	500	1000	1500	Baseline	0.5	1	3
0.1	92.3	88.8	90.3	91.1	91.2	91.0	91.0	53.0	81.0	87.1	87.1	75.9	80.1	85.0	81.7	81.9	82.4	60.9	67.6	69.4	69.4	
0.25	92.4	88.5	89.8	91.2	90.3	90.6	90.8	55.6	81.4	86.0	88.0	79.1	82.0	86.7	83.4	83.6	84.0	63.2	70.0	71.4	71.4	
0.5	93.6	90.2	91.3	92.9	91.3	91.5	91.2	56.8	83.0	86.6	89.0	80.0	82.9	87.5	84.3	84.4	84.7	64.9	71.3	73.1	73.1	
0.75	93.9	91.5	92.4	93.2	91.9	92.1	92.2	57.0	83.9	87.6	89.4	81.2	83.8	87.6	84.6	85.0	85.2	66.5	72.6	74.1	74.1	
1	94.5	91.8	92.8	93.9	92.3	92.5	92.6	57.7	84.1	87.6	89.4	81.6	83.9	87.8	84.9	85.2	85.3	66.4	72.9	74.4	74.4	

Table 1: Utility Results for *AG News* and *MNLI*. The scores represent the average micro-F1 score of the respective experiment over three runs. For readability, the standard deviations are not reported; however, any average with a standard deviation in the range (0.3, 0.8] is *italicized* (the highest recorded deviation was 0.8).

Split %	Baseline	Trustpilot										Yelp										
		1-DIFFRACTOR			DP-PROMPT			DP-BART				1-DIFFRACTOR			DP-PROMPT			DP-BART				
0.5	1	3	1.14	1.3	1.6	500	1000	1500	0.5	1	3	1.14	1.3	1.6	500	1000	1500	0.5	1	3	1.14	
0.1	Util ↑	99.3	97.0	98.4	98.5	98.1	98.5	98.5	94.0	97.8	98.6	97.4	94.6	94.7	96.8	85.1	89.4	91.7	82.1	89.1	91.6	91.6
	Priv (s) ↓	72.8	65.1	69.1	72.1	69.5	68.8	69.5	59.1	60.7	62.4	8.1	7.7	7.7	7.2	4.1	3.2	3.6	6.3	7.7	6.8	6.8
	Priv (a) ↓	72.8	64.5	62.4	67.9	68.1	68.3	67.3	59.5	62.0	64.5	8.1	5.0	6.2	6.6	3.5	3.6	4.4	4.7	5.7	7.7	7.7
	$\gamma$ (s) ↑	-	-0.68	-0.26	-0.26	-0.36	<b>0.28</b>	<b>-0.22</b>	-1.61	-0.34	<b>-0.10</b>	-	-0.33	-0.32	0.03	-1.18	-0.26	-0.22	-1.86	-1.08	-0.63	-0.63
	$\gamma$ (a) ↑	-	-0.67	-0.16	-0.20	-0.34	-0.21	<b>-0.20</b>	-1.62	-0.36	<b>-0.12</b>	-	0.00	-0.13	0.10	-1.11	-0.53	-0.32	-1.66	-0.83	-0.74	-0.74
	NN ↑	-	204	33	1	521	474	401	5229	1068	520	-	1	1	1	53	59	49	590	121	55	55
0.25	Util ↑	99.7	98.3	99.1	99.6	98.4	96.8	98.7	94.9	96.5	98.8	97.5	92.8	94.2	96.7	90.2	93.4	90.8	80.3	90.8	93.1	93.1
	Priv (s) ↓	73.9	67.2	70.0	73.5	69.7	69.5	70.6	61.0	62.1	62.5	22.5	12.7	14.5	18.5	8.5	7.1	7.6	6.9	5.8	8.2	8.2
	Priv (a) ↓	73.9	65.5	67.8	73.6	70.1	70.6	69.4	60.0	63.2	63.6	22.5	6.5	7.0	11.1	6.2	5.9	6.4	4.7	5.7	6.3	6.3
	$\gamma$ (s) ↑	-	<b>-0.33</b>	-0.13	-0.02	-0.33	-0.73	-0.26	<b>-1.27</b>	-0.80	-0.12	-	-0.09	-0.02	0.09	-0.20	<b>2.68</b>	-0.09	-1.24	-0.01	0.14	0.14
	$\gamma$ (a) ↑	-	<b>-0.31</b>	-0.10	-0.03	-0.34	-0.83	-0.24	<b>-1.26</b>	-0.82	-0.13	-	0.18	0.32	0.42	-0.10	<b>0.28</b>	-0.04	-1.14	-0.01	0.23	0.23
	NN ↑	-	347	58	2	744	658	577	6546	1662	855	-	2	1	1	147	138	129	1446	316	130	130
0.5	Util ↑	99.7	98.0	97.1	99.6	98.7	98.8	98.4	94.2	98.3	98.7	98.8	95.8	97.7	98.6	92.5	92.3	92.2	81.3	92.1	94.1	94.1
	Priv (s) ↓	74.7	66.9	69.7	73.5	70.6	71.2	71.1	60.9	60.7	60.9	45.6	14.3	16.7	29.6	14.1	15.5	15.7	7.3	9.6	8.2	8.2
	Priv (a) ↓	74.7	63.7	68.3	73.3	63.4	65.7	70.0	60.4	60.1	63.4	45.6	14.7	19.2	20.8	8.1	9.2	8.9	6.3	9.4	12.6	12.6
	$\gamma$ (s) ↑	-	-0.42	-0.74	<b>-0.01</b>	<b>-0.25</b>	0.05	-0.23	-0.152	-0.25	-0.12	-	0.40	0.53	0.33	0.09	0.61	0.03	-0.82	0.15	0.37	0.37
	$\gamma$ (a) ↑	-	-0.38	-0.72	-0.01	<b>-0.16</b>	<b>-0.19</b>	-0.22	-1.51	-0.24	-0.16	-	<b>0.39</b>	<b>0.47</b>	<b>0.52</b>	<b>0.22</b>	0.18	<b>0.18</b>	-0.80	0.16	0.28	0.28
	NN ↑	-	478	84	2	893	881	791	7660	2405	1358	-	3	1	1	309	270	257	2916	621	272	272
0.75	Util ↑	99.7	98.2	99.0	98.8	98.7	98.7	98.8	93.4	98.3	98.8	98.4	96.5	97.3	98.4	92.9	93.2	92.6	83.7	92.7	95.2	95.2
	Priv (s) ↓	74.4	66.8	69.8	73.8	70.6	70.9	71.0	61.4	61.9	62.2	43.7	8.8	11.8	22.6	12.5	12.4	13.2	6.4	7.4	7.6	7.6
	Priv (a) ↓	74.4	63.8	71.0	69.6	66.4	69.2	71.9	59.8	59.8	66.3	43.7	21.9	27.1	35.1	12.9	15.9	16.8	7.2	8.3	13.7	13.7
	$\gamma$ (s) ↑	-	-0.37	-0.16	-0.28	-0.27	0.05	-0.24	-1.82	-0.28	-0.12	-	0.60	0.62	0.48	<b>0.15</b>	0.79	0.10	<b>-0.66</b>	<b>0.25</b>	<b>0.50</b>	0.50
	$\gamma$ (a) ↑	-	-0.33	-0.18	-0.22	-0.21	-0.25	-1.80	-0.25	-0.18	-	0.30	0.27	0.20	0.14	0.10	0.02	<b>-0.67</b>	<b>0.22</b>	<b>0.36</b>	0.36	
	NN ↑	-	572	106	3	983	881	791	7681	2607	1508	-	5	1	1	486	441	418	4695	1143	516	516
1	Util ↑	99.7	96.1	99.2	99.6	98.7	98.8	98.7	94.4	98.4	98.7	98.5	96.6	97.6	98.4	92.7	92.6	93.1	83.2	92.7	94.9	94.9
	Priv (s) ↓	75.2	67.0	70.3	74.0	71.0	71.1	71.1	61.1	62.0	62.5	63.5	12.0	14.6	24.5	21.0	21.5	21.3	7.2	10.7	11.5	11.5
	Priv (a) ↓	75.2	63.5	63.7	69.3	69.4	69.9	70.4	60.1	62.1	70.4	63.5	29.3	44.9	52.8	19.6	20.7	21.4	8.6	17.1	21.9	21.9
	$\gamma$ (s) ↑	-	-0.99	<b>-0.09</b>	<b>-0.01</b>	<b>-0.25</b>	0.10	-0.25	-1.42	<b>-0.22</b>	-0.14	-	<b>0.62</b>	<b>0.68</b>	<b>0.60</b>	0.08	0.64	0.11	-0.67	0.24	0.45	0.45
	$\gamma$ (a) ↑	-	-0.94	<b>0.00</b>	<b>0.05</b>	-0.23	-0.20	-0.24	-1.41	<b>-0.22</b>	-0.24	-	0.34	0.20	0.16	0.10	0.07	<b>0.11</b>	-0.70	0.14	0.29	0.29
	NN ↑	-	641	123	3	1045	937	845	7881	2607	1508	-	5	1	1	486	441	418	4695	1143	516	516

deniability via nearest neighbors (*NN*). For each private dataset, we iteratively swap a private text with its original counterpart, then run a nearest-neighbor search to measure at which  $k$  the original text is the  $k$ -th nearest neighbor to the private text. A higher average  $k$  value would thus imply that the private counterparts are relatively indistinguishable from the originals. For performance reasons, we limit the top- $k$  search to 10,000; thus, a score of 10,000 would represent the best privacy.

All results are found in Table 1 for AG News and MNLI, and Table 2 for Trustpilot and Yelp. We also plot the average trade-off ( $\gamma$ ) value over all datasets, for each dataset split % in Figure 1.

**Reproducibility.** We note that for all experiments, a single-GPU server equipped with a 48GB Nvidia RTX A6000 GPU was utilized, along with 128GB of main memory. The entirety of the work required to complete the experiments described took approximately 2.5 months of GPU time on this

system. All training procedures use HuggingFace Trainer default parameters, and they are repeated three times for each dataset on different shuffles of the train split. For all sampling procedures throughout this work, namely dataset split preparation and train/val splitting, a random seed of 42 was used.

### 3.4 Scaling Up with Twitter

Beyond the Trustpilot and Yelp datasets, we sought to validate our findings on a larger-scale dataset, as the largest size dataset (Trustpilot) is limited to 360k reviews. To the best of the authors' knowledge, there does not exist a publicly available dataset that allows for the two-sided experiments we conduct in this work, namely with an associated utility and privacy task. Therefore, we constructed a new dataset based on an existing very large-scale corpus of Tweets (Enryu, 2023).

We decided to focus on the *authorship identification* adversarial task, as with the Yelp dataset.

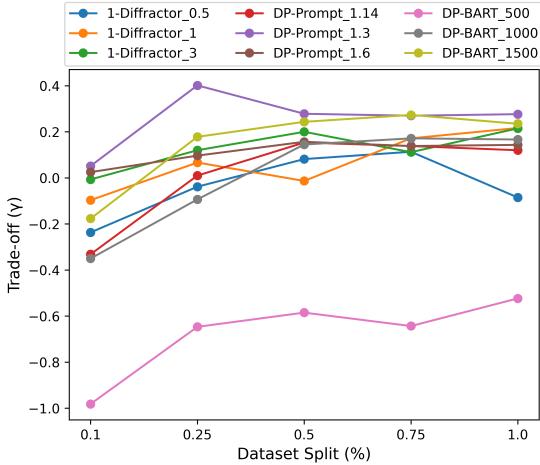


Figure 1: A visualization of average privacy-utility trade-offs ( $\gamma$ ) per split percentage. For each mechanism- $\varepsilon$  pair, we average all trade-offs (both static and adaptive, over all three utilized privacy evaluation datasets).

To construct a fitting dataset from the Twitter corpus, which contains around 88 million tweets, we first iterated through the corpus and enumerated the number of tweets authored by each unique user (by ID). Before this, however, we removed tweets detected not to be in English<sup>4</sup>. Then, we kept only the tweets from the top 1000 users, each of whom has 1001 tweets attributed to them (the maximum of any author in the dataset). This resulted in a dataset of just over one million tweets, which we give the title *Twitter\_1Kx1K*<sup>5</sup>.

To connect this dataset with a utility task, we assign a sentiment score to each tweet. We use the simple Vader SENTIMENTINTENSITYANALYZER from NLTK (Bird et al., 2009) to assign one of *positive* (1), *neutral* (0), or *negative* (-1) to a tweet. We use the *compound* score calculated by Vader, given on the scale of -1 to 1, and follow the guidelines of the original implementation<sup>6</sup>, namely scores above 0.5 as positive, below -0.5 as negative, and neutral for the rest. The resulting sentiments were distributed in roughly a 50/33/17 positive/neutral/negative split. Thus, we allow for a three-class sentiment classification task. The dataset is released with a CC-BY-4.0 license, mirroring the original corpus from Enryu (2023).

Following the dataset creation step, we followed the same privatization procedure with the Twitter data as was performed for Trustpilot and Yelp, using the three chosen DP mechanisms. Note that

<sup>4</sup>Using LINGUA-PY (<https://github.com/pemistahl/lingua-py>)

<sup>5</sup>The dataset is publicly available at [https://huggingface.co/datasets/sjmeis/Twitter\\_1Kx1K](https://huggingface.co/datasets/sjmeis/Twitter_1Kx1K)

<sup>6</sup><https://github.com/cjhutto/vaderSentiment>

in the case of DP-BART, we limit the dataset to only the top 100 authors (i.e., a 10% split), due to the considerable computation hours required to run DP-BART. We call this the *Twitter\_100x1K* subset. Following privatization, we followed the same training and metric calculation procedures as with Trustpilot and Yelp. The results are found in Table 3. Additionally, the detailed calculations for the  $\gamma$  metric is included as part of Appendix A.

## 4 Regression Analysis

Building on the results in Tables 2 and 3, we conduct a regression analysis to quantify the effect of dataset size on the DP text rewriting, along with other associated variables. As a target variable, i.e., the dependent variable, we choose to predict relative gain ( $\gamma$ ), as this gives an overall perspective on the effectiveness of DP text rewriting. In particular, we set the target to be the *average*  $\gamma$ , namely the average between the static and adaptive  $\gamma$  scores.

To predict the  $\gamma$  score, we define a number of dependent variables, which are based on our experimental setup and results:

- **Dataset size:** we use the size of the corresponding datasets in all of our experimental runs, considering the split size based on the *split %* of the overall dataset. Due to the very disparate range of dataset sizes, we take the natural logarithm of this value.
- **Average number of words:** we assume that an important quantity in the prediction of relative gains is not only the size of the dataset, but also the average text length within the dataset. As such, we calculate the average word length of each dataset split, using the NLTK.WORD\_TOKENIZE function.
- **Mechanism type:** we convert our three selected mechanisms into a categorical variable, specifically 1-DIFFRACTOR=1, DP-PROMPT=2, and DP-BART=3. While we realize that the intricacies of different DP mechanisms are much more complex, we justify this decision by considering the lexical level on which these mechanisms operate, i.e., ranging 1=*word-level* to 3=*document-level*.
- $\varepsilon$ : as with mechanism, we also convert the three tested  $\varepsilon$  values per mechanism into a categorical (1, 2, 3) variable, as the precise values are mechanism-specific and not directly comparable. As such, we interpret this categorical

Split %		Twitter						
		Baseline	1-DIFFRACTOR			DP-PROMPT		
			0.5	1	3	1.14	1.3	1.6
0.1	Util $\uparrow$	91.2	80.9	84.1	89.4	49.8	51.1	52.1
	Priv (s) $\downarrow$	2.8	2.0	2.1	2.5	0.2	0.1	0.5
	Priv (a) $\downarrow$	2.8	2.3	2.3	2.3	0.1	0.1	0.2
	$\gamma$ (s) $\uparrow$	-	0.19	0.17	0.11	0.49	<b>0.51</b>	0.40
	$\gamma$ (a) $\uparrow$	-	0.07	0.12	<b>0.18</b>	0.51	<b>0.52</b>	0.51
	NN $\uparrow$	-	4983	4986	4991	4970	4992	5002
0.25	Util $\uparrow$	95.1	84.2	87.4	93.4	50.4	50.7	51.8
	Priv (s) $\downarrow$	10.6	6.6	7.4	8.3	0.2	0.4	1.0
	Priv (a) $\downarrow$	10.6	8.7	8.6	9.8	0.1	0.1	0.3
	$\gamma$ (s) $\uparrow$	-	0.26	0.22	0.20	<b>0.51</b>	0.49	0.45
	$\gamma$ (a) $\uparrow$	-	0.06	0.11	0.06	<b>0.52</b>	<b>0.52</b>	0.51
	NN $\uparrow$	-	7989	7988	8009	7968	8010	7994
0.5	Util $\uparrow$	96.3	85.3	88.5	94.4	49.9	50.4	52.4
	Priv (s) $\downarrow$	26.8	15.4	16.5	18.4	0.3	0.5	1.4
	Priv (a) $\downarrow$	26.8	<i>18.6</i>	22.6	24.1	0.1	0.1	0.5
	$\gamma$ (s) $\uparrow$	-	0.31	0.30	0.29	<b>0.51</b>	0.50	0.49
	$\gamma$ (a) $\uparrow$	-	<b>0.19</b>	0.08	0.08	0.51	<b>0.52</b>	<b>0.52</b>
	NN $\uparrow$	-	8994	9000	9003	8987	9013	9003
0.75	Util $\uparrow$	96.9	85.7	89.1	95.0	50.2	50.5	52.2
	Priv (s) $\downarrow$	40.1	21.8	23.8	26.8	0.3	0.6	1.7
	Priv (a) $\downarrow$	40.1	29.8	<i>30.5</i>	32.2	0.1	0.1	<i>0.6</i>
	$\gamma$ (s) $\uparrow$	-	0.34	0.32	0.31	<b>0.51</b>	0.50	0.49
	$\gamma$ (a) $\uparrow$	-	0.14	<b>0.16</b>	<b>0.18</b>	0.51	0.51	<b>0.52</b>
	NN $\uparrow$	-	9330	9334	9338	9322	9341	9332
1	Util $\uparrow$	97.3	86.3	89.5	95.4	50.0	<i>50.6</i>	52.3
	Priv (s) $\downarrow$	45.8	23.9	26.0	29.2	0.3	0.6	1.8
	Priv (a) $\downarrow$	45.8	<i>36.0</i>	<i>34.8</i>	<i>38.0</i>	0.1	0.1	0.7
	$\gamma$ (s) $\uparrow$	-	<b>0.36</b>	<b>0.35</b>	<b>0.34</b>	<b>0.51</b>	0.50	<b>0.50</b>
	$\gamma$ (a) $\uparrow$	-	0.10	<b>0.16</b>	0.15	0.51	0.51	<b>0.52</b>
	NN $\uparrow$	-	9498	9502	9504	9495	9502	9502

Table 3: Utility and Privacy Results for *Twitter*. On the left-hand side are presented the results from 1-DIFFRACTOR and DP-PROMPT on the *Twitter\_1Kx1K* dataset, whereas the right-hand side presents the results from DP-BART on the smaller *Twitter\_100x1K* subset. The best relative gains ( $\gamma$ ) for each (*mechanism,  $\epsilon$* ) pair are **bolded**, for both the static (s) and adaptive (a) settings. *Italicized* results denote standard deviations above 0.3.

variable as *low*, *medium*, and *high* privacy budgets, or, from strictest to least strict privacy.

- **Utility labels:** this represents a simple count of the *support* of the label set of the associated utility task, as we presume this has some impact on the relative gain.
- **Privacy labels:** similarly, we include the support of the *privacy labels* for the associated (adversarial) privacy task. We take the natural logarithm of this value.

With these, we run a multivariate Ordinary Least Squares regression using the STATSMODELS library, with all default settings. Summary statistics of the fitted OLS model are provided in Table 4.

#### 4.1 Post-hoc Tests

Following the OLS regression, we conduct a deeper analysis of the results, asking the question of whether there exists any significant differences *between* dataset sizes, beyond the previous finding

$R^2 = 0.546$	coef.	std err	t	P> t
constant	-8.2040	1.282	-6.399	0.000
size (log)	0.1359	0.039	3.497	0.001
avg. # words	0.0126	0.002	5.775	0.000
mechanism	-0.1827	0.045	-4.039	0.000
$\epsilon$	0.1743	0.039	4.449	0.000
# util. labels	2.8812	0.454	6.350	0.000
# priv. labels (log)	-0.3209	0.072	-4.465	0.000

Table 4: MLR to predict the average  $\gamma$  score. In general,  $R^2$  measures the goodness of the fit, ranging from 0 to 1. All predictors are statistically significant.

that there exists *some* significant relationship between dataset size and relative gains.

To prepare this test, we first bin the (log) size variable into five equal bins calculated from the min and max observed values (using PANDAS.CUT), resulting in the following bins:

$$\{1: (7.698, 8.922] < 2: (8.922, 10.145] < 3: (10.145, 11.369] \\ < 4: (11.369, 12.593] < 5: (12.593, 13.817]\}$$

We convert these bins into a categorical variable, namely 1 for the lowest bin and 5 for the highest. Following this, we run a Kruskal-Wallis analysis

of variance (using `SCIPY.STATS`), which results in a p-value of 0.00087, thus allowing us to reject the null hypothesis that the population medians of all size groups are equal. Since we do not make any assumption of normality, we then proceed to run a Dunn’s post-hoc test (Dunn, 1964), which informs us of any significant differences between the categorical size variables. The results of this test are given in Table 5, which finds several significant pairwise relations between binned dataset sizes.

	1	2	3	4	5
1	1.0000	<b>0.0000</b>	0.1338	0.3060	<b>0.0148</b>
2	<b>0.0000</b>	1.0000	<b>0.0231</b>	<b>0.0001</b>	0.0593
3	0.1338	<b>0.0231</b>	1.0000	0.4332	0.4987
4	0.3060	<b>0.0001</b>	0.4332	1.0000	0.0660
5	<b>0.0148</b>	0.0593	0.4988	0.0660	1.0000

Table 5: Dunn’s post-hoc test results. **Bolded**  $p$ -values indicate statistically significant results ( $p < 0.05$ ).

## 5 Discussion

We reflect on the main findings of our experiments and discuss their implications for DP NLP.

**Dataset size is important.** Studying the results of our experiments, we find that dataset size is a very important factor when judging the efficacy of DP text rewriting, both from a utility and privacy standpoint. Looking to Table 2, one can see that given a single mechanism, the “best-performing” configuration, i.e., in terms of relative gains, can vary depending on dataset size and task. As an example, in the Yelp task, DP-BART only begins to exhibit positive relative gains at larger dataset split sizes, and generally speaking, all three tested mechanisms only begin to demonstrate positive relative gains at or above the 50% split mark.

Looking to the regression results, we find that dataset size is indeed a significant factor in predicting the expected relative gain for DP text rewriting. In particular, our fitted model suggests a relatively strong positive relationship between dataset size and relative gain, namely that with higher dataset sizes, we can expect average relative gains to rise. The post-hoc tests provide a deeper view into this significant relationship, highlighting further evidence of differences in observed trade-offs when the size of the data to be privatized varies.

Beyond this, an important finding lies in the fact that given different privacy budgets with the same mechanism, the most favorable trade-offs may not always be with the same dataset size. This sup-

ports the idea that more dynamic testing in terms of dataset size should be performed to provide a more holistic evaluation picture. Such testing is especially needed considering the task-specific differences that may be present (e.g., Trustpilot vs. Yelp), where the factors of adversarial task, dataset size, and mechanism effects can all interplay in a complex way. This is discussed next.

**What factors matter for favorable privacy-utility trade-offs?** Our regression analysis sheds light on this matter, as well as highlights remaining gaps in understanding. Importantly, we find that all included independent variables in our analysis prove to be significant to some degree in predicting relative gains from DP text rewriting. This not only includes dataset size, as discussed above, but also the *makeup* of the dataset, including average text length, the associated utility task, and even the nature of adversarial risk related to the dataset.

Beyond this, we find that the choice of DP mechanism is also a significant factor, yet the coefficient of *mechanism* implies that as we approach document-level methods (e.g., DP-BART), we might expect average relative gains to decrease. On the other hand, we observe from the empirical results that DP-BART on average experiences the largest *absolute* improvements as dataset size increases; however this comes with the caveat of often starting at a much lower point. We therefore observe the effect of document-level methods, where significant noise leads to a poor starting point, but with more data, may still exhibit utility. These are important findings, as the results suggest that mechanism specifics are significant in influencing expected trade-offs, providing motivation for novel methods that operate on diverse linguistic units.

An interesting insight comes with the regression results related to  $\epsilon$ , as the coefficient suggests that as  $\epsilon$  increases (i.e., the privacy guarantees become weaker), the average relative gains tend to increase. This may initially seem like a discouraging result for the field of DP text rewriting, yet one must consider that along this line, there exists some balance between chosen  $\epsilon$  and (reasonably) positive relative gain. Nevertheless, the true relationship between  $\epsilon$  and the privacy-utility trade-off most likely runs much deeper, and also is certainly intertwined with that of *mechanism* and other variables.

Despite the rich insights provided by conducting a regression analysis on the experiment results, we learn that there still exist gaps in understand-

ing what exactly influences the expected outcomes from DP text rewriting. This is signified by the  $R^2$  score of our fitted model (0.546), which suggests a good fit, yet one with a relatively high amount of variance still unexplained. Therefore, this implies that there still exist potentially numerous other factors important to the picture, which we simply did not include as variables in fitting the model. Additionally, some correlations (such as increasing  $\varepsilon$  leading to increasing  $\gamma$ ) must be tested more rigorously at scale. Finally, the effect of *task setup*, i.e., the number of utility and privacy “labels”, is only surfaced but merits further investigation. Such results provides motivation for future research into achieving a better understanding of the factors influencing local DP text rewriting outcomes.

**There is utility in DP.** A promising result, which can be extracted from both Tables 1 and 2, is that as dataset size grows larger, downstream utility of DP privatized text also generally increases. While this would generally be assumed outside of the DP context, these findings show promise that there is indeed utility to be gained from privatized text, and that these gains can become more apparent with larger dataset sizes. In particular, Table 1 demonstrates that as more data is available, DP text outpaces non-private text in “closing the gap” to the highest attainable utility. Albeit, this is due in part to the lower starting point of privatized data, yet the findings here show that using smaller dataset sizes may in fact be showing a *lower bound* of downstream utility for DP rewritten text.

**DP rewriting at scale? An outlook.** The above considerations lead to a very important question: can DP text rewriting, or more generally DP text privatization, make sense *at scale*? Such a discussion is crucial in weighing the potential benefits of DP text privatization in practical settings.

In favor of the affirmative answer to this question are several findings from this work. Firstly, we can observe that (significant) positive relative gains are possible in certain scenarios, particularly in those with larger dataset sizes. Beyond this, the *NN* metric clearly shows the *indistinguishability in a crowd* effect – the larger the dataset, the better protected a text becomes from its original. Finally, as dataset size grows, the typical DP utility hit becomes less apparent, and in some (*mechanism,  $\varepsilon$* ) scenarios, one can achieve utility quite close to the non-private baselines (see AG News or Trustpilot, for example). This result is supported by the signifi-

cance found in our regression model for increasing dataset size as a factor for higher relative gains. These findings shed light on the promise of DP text rewriting in practice and at scale.

However, there also come a few important considerations to this question. The first is a practical caveat: since the mechanisms tested in this work operate with *local* DP, having larger dataset sizes would imply a data processor who is capable of (and trusted to) process larger data volumes, as it is unlikely that single users would possess such data. Secondly, it is still unclear what is the *upper bound* of the effects observed in this work, and even larger-scale tests are required to continue investigating the effect of dataset size. For this, however, massive datasets would be needed, and furthermore, ones that can be reasonably attributed to some sensitive or adversarial scenario. This becomes crucial going forward, i.e., establishing high-quality and practical benchmarks for testing DP text privatization.

On this note, our findings point to an important consideration in the design of privacy benchmarking DP text rewriting. In our regression analysis, we find that the nature of both the utility and privacy tasks (as proxied by the number of associated labels) significantly impacts the quantification of relative gain. While further tests on a wider variety of downstream tasks would be needed to validate the regression coefficients, we nevertheless learn of the influence that evaluation setup decisions may have on the measurement (and eventual perception) of the effectiveness of DP text rewriting.

## 6 Conclusion

We investigate the effect of dataset size in DP text rewriting, using five datasets to evaluate both utility and privacy in various rewriting scenarios. Our results suggest that while larger dataset sizes are not a silver bullet for effective text privatization, they generally lead to more favorable trade-offs. This, however, is mechanism- and task-specific, showcasing the complexity of local DP text rewriting, and of DP NLP in general. Furthermore, we empirically demonstrate the importance of varying dataset size in evaluation procedures, showing that measured trade-offs may differ based on the size of the utilized datasets. In this, we provide an outlook for DP text privatization and evaluation at scale, proposing that with further (large-scale) testing, proposed techniques from the literature may begin to realize their potential in practice.

## Acknowledgments

The authors would like to thank the reviewers for their helpful feedback in improving this work, as well as Alexandra Klymenko for her valuable support in preparing the paper.

## Limitations

A primary limitation with our work comes with the definition of “large-scale” datasets. While we utilized datasets that were known to us, publicly available, and reasonable to use within the confines of our resources, we are aware that the chosen datasets may still be considered small in the light of “big data”. However, we posit that the findings resulting from our experiments provide a starting point for even larger-scale experiments, especially in comparison to previous DP NLP works on smaller dataset samples.

Another limitation is one common to many related works: the evaluation of *privacy*. We chose a combination of attribute-based empirical privacy (mitigation of inference attacks), as well as a designed indistinguishability test. We choose these in order to follow the evaluation procedure of previous works, especially in light of no standard benchmark for DP text rewriting. As such, the results we provide are a *proxy* for privacy. Our relative gain ( $\gamma$ ) calculations are therefore based on this proxy.

Finally, as showcased by our results, the effect of DP rewriting at scale can be significantly different from one mechanism to another. Our choice of inherently different rewriting mechanisms is a testament to that fact, yet we do not analyze in detail the potential technical origins of these differences, for example, in light of word- versus document-level rewriting. Future works could supplement our findings with such an analysis, particularly by including more than our chosen three mechanisms.

## Ethics Statement

As our work is focused in the field of privacy-preserving NLP, it places itself directly parallel to ethical and responsible NLP and AI. We hope that with the findings of our work, we may contribute to the body of knowledge of how to conduct (and evaluate) NLP in a privacy-preserving manner.

One ethical consideration comes with the simulation of adversaries in our privacy experiments. This is performed using publicly available datasets not originally intended for these adversarial purposes. Nevertheless, such considerations are mitigated

since the datasets used (1) are long-standing in the research sphere, (2) contain no PII, and (3) have been used similarly by previous DP NLP works.

## References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. *Deep learning with differential privacy*. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Stefan Arnold. 2025. *Inspecting the representation manifold of differentially-private text*. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 53–59, Albuquerque, New Mexico. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O’Reilly Media, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. *Scaling instruction-finetuned language models*. *arXiv preprint*.

Olive Jean Dunn. 1964. *Multiple comparisons using rank sums*. *Technometrics*, 6(3):241–252.

Cynthia Dwork. 2006. *Differential privacy*. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Enryu. 2023. Dataset of 100m tweets, see also [https://huggingface.co/datasets/enryu43/twitter100m\\_tweets](https://huggingface.co/datasets/enryu43/twitter100m_tweets).

Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021. *Research challenges in designing differentially private text generation mechanisms*. In *The International FLAIRS Conference Proceedings*, volume 34.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. *Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations*. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM ’20, page 178–186, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. *Deberta: Decoding-enhanced bert with disentangled attention*. In *International Conference on Learning Representations*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. [User review sites as a resource for large-scale sociolinguistic studies](#). In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. [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.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. [DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. [DP-BART for privatized text rewriting under local differential privacy](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2008. [What can we learn privately?](#) In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 531–540.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. [Differential privacy in natural language processing the story so far](#). In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. [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.

Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. [Sentence-level privacy for document embeddings](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3367–3380, Dublin, Ireland. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024a. [1-Diffractor: Efficient and utility-preserving text obfuscation leveraging word-level metric differential privacy](#). In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics, IWSPA '24*, page 23–33, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025. [On the impact of noise in differentially private text rewriting](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 514–532, Albuquerque, New Mexico. Association for Computational Linguistics.

Stephen Meisenbacher and Florian Matthes. 2024. [Thinking outside of the differential privacy box: A case study in text privatization with language model prompting](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5656–5665, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024b. [A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 174–185, Torino, Italia. ELRA and ICCL.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. [How to dp-fy ml: A practical guide to machine learning with differential privacy](#). *Journal of Artificial Intelligence Research*, 77:1113–1201.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. [Locally differentially private document generation using zero shot prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. 2024. [Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 507–527, Miami, Florida, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. [DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders](#). In *Proceedings of the ACM Web Conference 2022*, WWW ’22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, page 649–657, Cambridge, MA, USA. MIT Press.

## A Relative Gain Calculation

For the Trustpilot dataset, the complete dataset is heavily imbalanced towards positive reviews (93.07%), yielding an overall  $MG_u$  of 96.41%. Accordingly, for the val set on all dataset splits, we calculate:

- **10%**:  $MG_u = 96.36\%$
- **25%**:  $MG_u = 96.38\%$
- **50%**:  $MG_u = 96.47\%$
- **75%**:  $MG_u = 96.54\%$
- **100%**:  $MG_u = 96.41\%$

Similarly, for the Yelp dataset, the reviews are primarily positive (78.74%), for an overall  $MG_u$  of 88.11%. This yields the following values, calculated from the 10% val split:

- **10%**:  $MG_u = 90.05\%$
- **25%**:  $MG_u = 88.60\%$
- **50%**:  $MG_u = 88.30\%$
- **75%**:  $MG_u = 88.66\%$
- **100%**:  $MG_u = 88.69\%$

Finally, for the Twitter dataset, the tweets are primarily positive (49.48%), for an overall  $MG_u$  of 66.20%. This yields the following values, calculated from the 10% val split:

- **10%**:  $MG_u = 67.07\%$

- **25%**:  $MG_u = 67.06\%$
- **50%**:  $MG_u = 66.95\%$
- **75%**:  $MG_u = 66.86\%$
- **100%**:  $MG_u = 67.14\%$

We also calculate the values for the smaller sample used for DP-BART:

- **10%**:  $MG_u = 65.99\%$
- **25%**:  $MG_u = 68.17\%$
- **50%**:  $MG_u = 67.16\%$
- **75%**:  $MG_u = 67.42\%$
- **100%**:  $MG_u = 67.27\%$

These values are used for their respective dataset splits in the  $\gamma$  calculations for Tables 2 and 3.