A Comparative Analysis of Word-Level Metric Differential Privacy: Benchmarking The Privacy-Utility Trade-off

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, Florian Matthes
Technical University of Munich
TUM School of Computation, Information and Technology
Department of Computer Science
Garching, Germany
{stephen.meisenbacher,nihildev.nandakumar,alexandra.klymenko,matthes}@tum.de

Abstract

The application of Differential Privacy to Natural Language Processing techniques has emerged in relevance in recent years, with an increasing number of studies published in established NLP outlets. In particular, the adaptation of Differential Privacy for use in NLP tasks has first focused on the *word-level*, where calibrated noise is added to word embedding vectors to achieve "noisy" representations. To this end, several implementations have appeared in the literature, each presenting an alternative method of achieving word-level Differential Privacy. Although each of these includes its own evaluation, no comparative analysis has been performed to investigate the performance of such methods relative to each other. In this work, we conduct such an analysis, comparing seven different algorithms on two NLP tasks with varying hyperparameters, including the *epsilon (ε)* parameter, or privacy budget. In addition, we provide an in-depth analysis of the results with a focus on the privacy-utility trade-off, as well as open-source our implementation code for further reproduction. As a result of our analysis, we give insight into the benefits and challenges of word-level Differential Privacy, and accordingly, we suggest concrete steps forward for the research field.

Keywords: differential privacy, privacy-preserving NLP, evaluation

1. Introduction

Privacy vulnerabilities in Natural Language Processing (NLP) have recently been placed in the spotlight, and the discussions surrounding data privacy in this setting have gained increased attention with the rise of Large Language Models (LLMs) and chatbots such as ChatGPT. In particular, privacy risks have been demonstrated in embedding models (Song and Raghunathan, 2020; Thomas et al., 2020; Morris et al., 2023) and general-purpose language models (Pan et al., 2020; Carlini et al., 2021).

To combat privacy risks in data processing settings, Privacy-Enhancing Technologies (PETs) have emerged as concrete technical solutions that can be incorporated into existing systems. Under this class of technologies, Differential Privacy (DP) (Dwork, 2006) has risen in popularity due to its mathematical foundations, composability and robustness to post-processing, and above all, its flexible privacy parameter, known as ε.

The application of DP to NLP settings does not come immediately, as the original sense of DP was designed for injecting plausible deniability into queries performed on sensitive attributes from structured databases. As textual data rarely exists in this form, reasoning about DP definitions initially comes with its challenges (Klymenko et al., 2022). Nevertheless, a number of implementations have appeared in the literature, and as pointed out by Hu et al., the majority of these revolve around embedding vector perturbation methods at the word level (Hu et al., 2023). Many of these implementations employ *Metric* Local Differential Privacy (MLDP), which was introduced as a generalization of the standard DP notion (Chatzikokolakis et al., 2013).

The focus on applying DP to word embeddings marks an intuitive first step in fusing the two fields, as words can be perceived as atomic units of information, which in turn are replaceable via calibrated perturbations. In such methods, the goal becomes to obfuscate the original text data as much as possible, while still preserving semantic coherence, and ideally, grammatical correctness. In terms of privacy preservation, several metrics are introduced in the literature, such as plausible deniability statistics (Feyisetan et al., 2020) or membership inference attack performance (Shokri et al., 2017; Carvalho et al., 2023). Even so, such statistics are not uniformly reported across all word-level DP papers.

Beyond the metrics used to quantify the implications on privacy and utility, implementation papers do not run a standard evaluation, making a comparison in terms of performance quite difficult. The diversity in evaluation setups can be attributed both to the relative adolescence of the field and accordingly, the lack of a defined benchmark.

In this work, we aim to address some of the above-mentioned gaps. We design an experimental setup with two separate NLP tasks, in which seven different word-level DP algorithms are tested.

LREC-COLING 2024, pages 174–185
20-25 May, 2024.

These experiments are run with various combinations of ε and embedding dimension. Finally, a set of statistics is calculated on each experiment iteration, providing the foundation for a comparative analysis against the provided baselines.

The results from this work present the following contributions to the research of DP in NLP:

1. An overview of the disparity in evaluation methods for word-level DP
2. A novel multi-dimensional experimental setup focused on benchmarking privacy and utility metrics for word-level MLDP
3. A comparative analysis of word-level MLDP methods, guided by a novel composite metric
4. An open-source replication package for reproduction of the experiments, which includes previously unavailable code implementations of the selected methods, found at: https://github.com/sjmeis/MLDP

The structure of this paper is as follows. In Section 2, related work in the field of word-level DP and its evaluation are discussed. Afterwards, in Section 3, foundations of DP for NLP are introduced. Section 4 briefly outlines the followed methodology for this work, while Section 5 illustrates the resulting findings. These results are analyzed and discussed in Section 6. Finally, Section 7 underlines the implications following from our work and potential future directions, which is followed by a discussion of the perceived limitations of our study.

2. Related Work

The investigation of Differential Privacy in Natural Language Processing, specifically on the word level, can be traced back to SynTF (Weggenmann and Kerschbaum, 2018), in which "synthetic" term-frequency vectors are created by performing single word replacements using the Exponential Mechanism (McSherry and Talwar, 2007). Fernandes presented the novel concept of using calibrated noise added directly to word embedding vectors to achieve noisy, perturbed vectors (Fernandes et al., 2019). This method relies on a generalized form of DP, often referred to as *metric* DP, which relaxes DP for use in arbitrary vector spaces endowed with a metric (Chatzikokolakis et al., 2013). Further improvements to this technique were achieved by experimentation with underlying noise addition mechanisms, distance metrics, or both (Xu et al., 2020; Feyisetan et al., 2020; Carvalho et al., 2023). These implementations focus on the *local* DP setting (Kasiviswanathan et al., 2011), in which DP is applied to data at the user level and not at some central authority.

A recent survey (Hu et al., 2023) categorizes DP-NLP methods into two categories: *gradient perturbation* and *vector embedding perturbation*. Of the 19 implementations listed under vector embedding perturbation in the local setting, 17 are word-level.

Klymenko et al. (2022) highlight the importance of benchmarking in DP-NLP, particularly as future research in the field. Looking to the word-level methods outlined by Hu et al., there is a great disparity in the tasks, datasets, and parameters used to evaluate the proposed methods. An overview of these evaluations, in line with the 17 mentioned methods, is provided in Table 1.

In the works presented in Table 1, utility is often measured by evaluating the accuracy of a given NLP task with perturbed input data. Privacy, on the other hand, is largely measured via (1) *Empirical Privacy*, or the *decrease* in performance for adversarial attacks, or (2) *Plausible Deniability*, in which statistics try to illustrate the level of plausible deniability introduced by a DP mechanism. Concretely, plausible deniability is often measured by estimating the probability that a word will be perturbed to another word, i.e., not remain the same.

Of particular focus in this work are the publications presented in the bottom half of Table 1. Specifically, we investigate MDP techniques in the local setting, henceforth Metric-LDP, or *MLDP*. These methods generally focus on leveraging MLDP to add calibrated noise to static word embeddings (e.g., GloVe (Pennington et al., 2014)), in order to achieve noisy word representations.

In another recent work (Mattern et al., 2022), the authors address potential shortcomings of word-level DP, particularly the tight constraints placed in the local DP setting, as well as the effect this has on the quality of language output, leading to grammatical errors and inflexibility when attempting to enforce changes in syntax. The implications of these findings on model performance were not discussed or analyzed, however, and it is here where our investigation begins.

3. Foundations

Differential Privacy (Dwork, 2006) was introduced in 2006 as a formal definition for the quantification of individual privacy. The original notion as proposed by Dwork was aimed at privacy preservation in the centralized setting, in which each row of a structured database corresponds to one individual's data. According to Differential Privacy, the inclusion of an individual in the dataset should only affect the outcome of aggregate queries by a certain bound, governed by the ε (privacy) parameter.

Definition 3.1 (ε-Differential Privacy)**.** For any databases D_1 and D_2 differing in exactly one element, any $\varepsilon > 0$, a randomized function $\mathcal{K}$, and all

Publication	Model	Task	Dataset	Epsilon	Privacy Metrics	Utility Metrics
(Lyu et al., 2020a)	BERT	Sentiment Analysis Topic Classification	Trustpilot AG News/DW	$\varepsilon \in \{0.05, 0.1, 0.5, 1, 5\}$	Empirical Privacy	-
(Lyu et al., 2020b)	GloVe BERT	Sentiment Analysis Intent Detection Paraphrase Identification	IMDb, Yelp, Amazon Intent Dataset MRPC	$\varepsilon \in \{0.5, 1, 5, 10\}$	-	Accuracy, F1
(Plant et al., 2021)	BERT	Sentiment Analysis	Trustpilot	$\varepsilon \in \{0.01, 0.1, 0.5, 1\}$	Empirical Privacy	Accuracy, F1
(Krishna et al., 2021)	LSTM	Intent Classification	ATIS, SNIPS	$\varepsilon \in \{0.25, 0.5, 0.6, 0.75, 0.85, 1\}$	AUC	Accuracy
(Habernal, 2021)	-	-	-	-	-	-
(Igamberdiev et al., 2022)	LSTM	Intent Classification	ATIS, SNIPS	$\varepsilon \in \{1, 10, 100, 1000\}$	-	F1
(Maheshwari et al., 2022)	Private Encoder	Sentiment Analysis Attribute Detection	Twitter Bias in Bios, CelebA, Adult Income	$\varepsilon \in \{8, 10, 12, 14, 16\}$	Empirical Privacy, MDL	Accuracy
(Feyisetan et al., 2020)	GloVE FastText	Binary Classification Multi-class Classification Question Answering	IMDb Enron emails InsuranceQA	$\varepsilon \in \{6, 12, 17, 23, 29, 35, 41, 47, 52\}$	Plausible Deniability	Accuracy
(Xu et al., 2020)	FastText	Binary Classification Word Classification	Twitter, SMSSpam Product Reviews	$\varepsilon \in \{1, 5, 10, 20, 40\}$	Plausible Deniability	Accuracy
(Xu et al., 2021b)	GloVe FastText	Sentiment Analysis Binary Classification	IMDb Twitter	$\varepsilon \in (0, 40]$	Empirical Privacy	-
(Xu et al., 2021a)	GloVe	Sentiment Analysis Textual Entailment	IMDb SNLI	N.A.	Empirical Privacy	Accuracy
(Carvalho et al., 2023)	GloVe	Sentiment Analysis	IMDb	$\varepsilon \in (0, 22]$	Empirical Privacy	Accuracy
(Feyisetan and Kasiviswanathan, 2021)	GloVe FastText	Various	MR, CR, MPQA, SST-5, TREC-6	N.A.	-	Accuracy
(Feyisetan et al., 2019)	Poincaré	Various	MR, CR, MPQA, SST-5, TREC-6 SICK-E, MRPC, STS14	$\varepsilon \in \{0.125, 0.5, 1, 2, 8\}$	Plausible Deniability	Accuracy
(Carvalho et al., 2021)	Binary	Sentiment Analysis	IMDb	$\varepsilon \in (0, 22]$	Empirical Privacy	Accuracy
(Tang et al., 2020)	GloVe	Sentiment Analysis Topic Classification	Trustpilot AG News	$\varepsilon \in \{3, 4, 5, 6, 7, 8\}$	Plausible Deniability	Accuracy
(Yue et al., 2021)	GloVe BERT	Sentiment Analysis Semantic Textual Similarity Question Answering	SST-2 MED-STS QNLI	$\varepsilon \in \{1, 2, 3\}$	Empirical Privacy	Accuracy

Table 1: Word-level Local Differential Privacy (LDP) techniques and their evaluation. The bottom section denotes word-level *metric* LDP approaches, the majority of which operate on static word embeddings.

$\mathcal{S} \subseteq Range(\mathcal{K})$:

$$Pr[\mathcal{K}(D_1) \in \mathcal{S}] \leq e^{\varepsilon} Pr[\mathcal{K}(D_2) \in \mathcal{S}] \quad (1)$$

Thus, the ε parameter determines how *indistinguishable* the output (distribution) of the operation performed on D_1 and D_2 must be.

As mentioned, the focus of this work is placed on the application of DP in the word embedding space, which is a multi-dimensional *vector space*. As such, the definition of Equation 1 is not readily transferable to this space. Instead, the notion of MDP was developed to incorporate the usage of a distance metric within the word vector space.

Definition 3.2 (Metric Differential Privacy, or $d_{\mathcal{X}}$-privacy)**.** For any $x, x' \in \mathcal{X}$ (vector space) endowed with a metric d, any $\varepsilon > 0$, and a randomized function $M : \mathcal{X} \to \mathcal{Y}$:

$$Pr[M(x) \in \mathcal{Y}] \leq e^{\varepsilon d(x,x')} Pr[M(x') \in \mathcal{Y}] \quad (2)$$

One can see that this definition incorporates the metric d into the ε parameter, now scaling the required indistinguishability by the relationship between two inputs (i.e., two words from the vocabulary $\mathcal{X}$). In this setting, one can now reason about two word vectors, whose relation can be quantified by a distance metric.

In order to define MDP in the local setting, the notion of MLDP was introduced (Alvim et al., 2018), which is defined below:

Definition 3.3 (Metric Local Differential Privacy)**.** For all $y \in \mathcal{Y}$:

$$Pr[M(x) = y] \leq e^{\varepsilon d(x,x')} Pr[M(x') = y] \quad (3)$$

For an in-depth introduction of DP in metric spaces for NLP, we refer the reader to (Feyisetan et al., 2020). For technical details on how calibrated noise is generated using a variety of DP mechanisms, we refer to (Barthe et al., 2016).

4. Methodology

In this section, we introduce the details of our experimental design, as well as provide a brief overview of the algorithms included in our analysis.

4.1. Experimental Design

4.1.1. Tasks and Datasets

For our comparative analysis, we benchmark the selected methods on two NLP tasks: Sentiment Analysis and Topic Classification.

The Sentiment Analysis task is run on the IMDb Movie Review Dataset (Maas et al., 2011), which is a dataset of 50k movie reviews, classified as either negative or positive in sentiment. We take a random sample of 12k movies – 8k for training, 2k for validation, and 2k for testing.

AG News (Zhang et al., 2015) is a dataset of nearly 130k text excerpts from AG. The dataset contains news articles on the four largest topics in the AG News corpus: world, sports, business, and science. For this task, we take a random sample of 6k articles from each topic – 16k in total for training, 4k for validation, and 4k for testing.

A summary of both selected tasks and their underlying datasets is included in Table 2.

Dataset	IMDb	AG News
Task Type	Binary	Multi-class
Training set size	10,000	20,000
Test set size	2,000	4,000
Total word count	808,382	510,582
Vocabulary Size	42,662	27,234
Sentence Length	μ = 80.84 σ = 22.53	μ = 25.52 σ = 6.78

Table 2: Summary of the two selected NLP tasks and datasets, with key characteristics.

Choice of Datasets We choose IMDb and AG News for multiple reasons: (1) their utilization in previous works (Tang et al., 2020; Carvalho et al., 2023; Xu et al., 2021b; Feyisetan et al., 2020), (2) their relatively large size and accessibility, (3) the ability of IMDb to simulate "sensitive" information (personal reviews), and (4) the multi-class classification problem of AG News, to supplement the simpler binary case of IMDb.

4.1.2. Evaluation Model

For both tasks, we employ an LSTM-based model (Hochreiter and Schmidhuber, 1997) using Keras. An embedding layer is added to allow the use of GloVe embeddings as input, followed by a Dropout layer (0.2), and finally, a fully connected layer with a softmax activation is added to facilitate both classification tasks. The embedding layer was added to facilitate the input format, as all tested mechanisms map input words to discrete "noisy" output words, each of which corresponds to an embedding (GloVe in this case). For each experiment configuration, the model was trained with a batch size of 64 with a maximum of 30 epochs. Early stopping and checkpointing were activated.

It should be noted that the choice of LSTM in lieu of transformer-based models was justified for two reasons: (1) as the nature of this study is to benchmark against a baseline, it was not seen as necessary to achieve SOTA performance on the two chosen tasks, and relatedly, (2) the training of LSTMs is far more efficient than that of transformer-based models, and given the large dimensions of our conducted evaluation, the LSTM provided the much more time- and resource-efficient option.

For each experimental setting, the MLDP perturbed data (train + validation split) is used for training, and the evaluation is performed on the trained model using the test split perturbed by the same mechanism. This is to simulate the local DP setting, in which user data is perturbed locally. The metrics of the models trained on perturbed data were captured and compared against the non-DP (original data) baseline.

4.1.3. Embedding Model

We utilize pre-trained GloVe embeddings (Pennington et al., 2014), which were trained on the entire Wikipedia dataset from 2014 and Gigaword 5[1]. In particular, we use the 50-, 100-, and 300-dimension versions of the word embedding model.

4.1.4. Privacy Budget

As introduced in Section 3, the privacy parameter known as epsilon (ε) is used to scale the level of privacy desired. In practice, a smaller epsilon value adds a higher level of noise and thus leads to a higher theoretical level of privacy protection. As a result of surveying the parameter choice of the MLDP methods in Table 1, we choose to use $\varepsilon \in \{1, 5, 10\}$ for our experiments.

4.1.5. Metrics

For utility, we report the accuracy of each experiment. This is seen to be a sufficient indicator of performance, as both tasks have balanced datasets.

For privacy estimation, we employ four metrics which allow for a uniform platform for evaluating the effect of each algorithm on the input text. The first three methods, introduced below, were chosen due to their usage in the previous works listed in Table 1, while the final metric (LOW) is inspired by the approach of Yue et al..

Plausible Deniability (PD) [$N_w \downarrow, S_w \uparrow$] These two metrics measure the *plausible deniability* created by a particular mechanism. Concretely, given a word n, N_w estimates the probability of not modifying the word (word stays unperturbed), and S_w measures the support of the set of output words, i.e., the number of words that w can be perturbed with probability $1 - \eta$, with η being small. Inspired by the approach taken by (Feyisetan et al., 2020; Xu et al., 2021b), we estimate these metrics by running each mechanism on a set of 25 random words from each dataset, each word being run 100 times.

Perturbation Percentage (PP) $\uparrow$ This metric calculates the percentage of words perturbed to a different word, i.e., that did not remain unchanged. The more perturbed the data is, the higher the privacy protection; however, this may come at the expense of reduced data utility (Kang et al., 2020).

Cosine Similarity (CS) $\uparrow$ This metric calculates the cosine similarity between the sentence embedding representations of the original and perturbed inputs, inspired by BERTScore (Zhang et al., 2019).

[1] https://nlp.stanford.edu/projects/glove/

Although not a privacy metric per se, cosine similarity provides insight into the effect of word-level perturbations. By using this metric in combination with PP, one can gauge the trade-off between input perturbation and preservation of meaning. To calculate this metric, a pre-trained SBERT model (Reimers and Gurevych, 2019) is used, namely **paraphrase-distilroberta-base-v1**.

Least-Occuring Words (LOW) $\downarrow$ This metric calculates the percentage of least-occurring words existing in both the original and perturbed datasets. Least-occurring words are considered sensitive because they may contain information about individuals, potentially leading to privacy breaches (Yue et al., 2021). Here, we calculate the percentage of the 1000 least-occurring words from the original dataset that are still present in the perturbed data.

4.2. Selected Algorithms

We introduce the seven methods included in this study, which represent all existing word-level MLDP approaches that operate on static word embeddings in the Euclidean space. In addition, we also include one method (SynTF) as a baseline, as it served as a precursor to all other evaluated methods. Finally, we also briefly discuss the excluded methods from Table 1.

SynTF (Weggenmann and Kerschbaum, 2018) Although not explicitly an MLDP method, the SynTF mechanism can be viewed as a precursor, as it performs word-level DP synonym replacements by sampling words from term-frequency vectors.

Calibrated Multivariate Perturbations (CMP) (Feyisetan et al., 2020) This method adds calibrated multivariate normal noise to word embeddings, and then perturbs the noisy vectors back to the nearest neighbor in the embedding space.

Mahalanobis Mechanism (Xu et al., 2020) This method aims to improve upon previous methods by adding elliptical noise using the regularized Mahalanobis norm, in order to account for vectors existing in sparse regions of the embedding space.

SanText (Yue et al., 2021) The SanText mechanism aims to improve word perturbation by relating the perturbation probability of a word to another token by their Euclidean distance in the embedding space. Thus, the closer two words are semantically, the higher the probability that one serves as the replacement for the other. We use the base mechanism proposed in the paper.

Truncated Gumbel Mechanism (Xu et al., 2021a) This method utilizes calibrated Gumbel noise to scale the probability of perturbing to a new word within a selected set of candidate words.

Vickrey Mechanism (Xu et al., 2021b) This mechanism, motivated by Vickrey auctions, balances the perturbation probability between the first and second nearest word neighbors. The authors also provide a generalized mechanism for k neighbors; we implement the original method (k=2).

Truncated Exponential Mechanism (TEM) (Carvalho et al., 2023) This mechanism generalizes the perturbation process to a *selection problem* by utilizing the Exponential Mechanism. For our study, we utilize the TEM mechanism with Euclidean distance, as proposed in the paper.

4.2.1. Excluded Algorithms

Of the introduced MLDP methods presented in Table 1, we exclude (Feyisetan et al., 2019) and (Carvalho et al., 2021) due to their use of embeddings in non-euclidean spaces. Similarly, we exclude (Feyisetan and Kasiviswanathan, 2021), as this method does not map noisy vectors to words. (Tang et al., 2020) is excluded due to its multi-stage perturbation mechanism and its similarity to CMP.

4.2.2. Algorithm-specific Parameters

In our experiments, we used the following algorithm-specific parameters (beyond ϵ), which can be found and modified in our provided repository: **SynTF**: synonyms from NLTK WordNet, **Mahalanobis**: $\lambda = 0.2$, **Vickrey**: $t = 0.5$, **TEM**: $\gamma = 0.5$.

5. Experiment Results

5.1. Utility

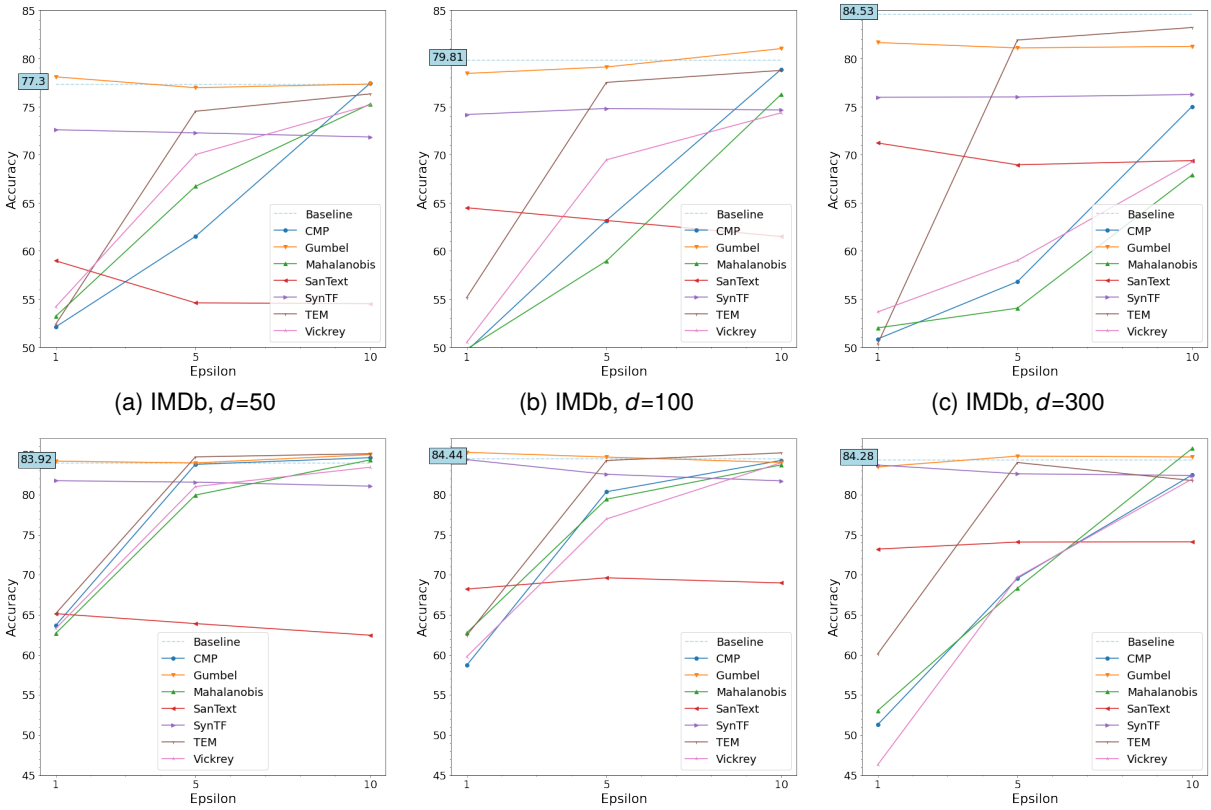
The utility results for our study are presented in Table 3. To obtain the accuracy scores, the LSTM model was trained five times (10 for baseline tests), and the evaluated scores were averaged to achieve a single score. As early stopping was implemented, accuracy was drawn from the best model. For each combination of (*task*, *dimension*, *epsilon*), the highest score is **bolded**. Scores that surpass their respective (*task*, *dimension*, *epsilon*) benchmark are underlined. The measured scores are broken down by embedding dimension and task in Figure 1.

5.2. Privacy

For readability, we present summarizing visualizations of the privacy metric results. Figure 2 illustrates the N_w:S_w ratio of the selected methods for

Task:	Sentiment Analysis (IMDb)									Topic Classification (AG News)								
Baseline:	77.30			79.81			84.53			83.92			84.44			84.28		
Dimension:	50			100			300			50			100			300		
Epsilon:	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
SynTF	72.58	72.26	71.83	74.16	74.79	74.64	75.96	75.99	76.24	81.73	81.55	81.05	84.36	82.53	81.71	**83.69**	82.59	82.39
CMP	52.10	61.50	**77.40**	49.60	63.15	78.85	50.85	56.80	75.00	63.67	83.77	84.60	58.70	80.35	84.26	51.30	69.55	82.45
Mahalanobis	53.20	66.70	75.25	49.80	58.95	76.25	52.00	54.05	67.90	62.67	79.94	84.32	62.75	79.42	83.70	53.05	68.32	**85.75**
SanText	58.97	54.60	54.52	64.48	63.16	61.49	71.21	68.94	69.38	65.15	63.91	62.46	68.21	69.62	68.98	73.20	74.08	74.11
Gumbel	**78.08**	**76.95**	77.33	**78.43**	**79.10**	**81.01**	**81.64**	81.08	81.24	**84.17**	83.95	84.98	**85.27**	**84.68**	84.00	83.45	**84.80**	84.69
Vickrey	54.18	70.00	75.16	50.50	69.46	74.34	53.66	59.00	69.23	63.29	81.00	83.41	59.78	76.95	83.97	46.29	69.70	81.94
TEM	52.38	74.50	76.30	55.15	77.50	78.75	50.34	**81.90**	**83.20**	65.15	**84.70**	**85.10**	62.42	84.25	**85.20**	60.12	84.00	81.75

Table 3: Utility scores (accuracy) for all experimental settings. The scores represent an average of five runs (10 for baseline). Bolded scores denote the highest score per setting, while underlined scores mark those that surpass the baseline.



(a) IMDb, *d*=50 (b) IMDb, *d*=100 (c) IMDb, *d*=300

(d) AG News, *d*=50 (e) AG News, *d*=100 (f) AG News, *d*=300

Figure 1: Accuracy scores per task and embedding dimension (*d*). Baseline scores are marked with a dotted line, and the baseline value is indicated in the light blue box. The scale of the y-axis is uniform across sub-figures for comparability.

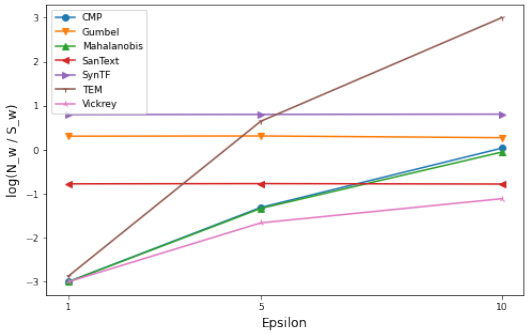
our three chosen ϵ values. As lower N_w and higher S_w values are preferred, lower values on the graph represent higher plausible deniability guarantees.

A New Composite Metric The study of text privatization, for example in the case of word-level MLDP, often views the privacy-utility trade-off in two separate lights: privacy and utility are measured separately, and then these results are fused in a qualitative analysis. As such, to the best of the authors' knowledge, there exists no single metric that compares privacy and utility *simultaneously*. We aim to address this gap in the introduction of the following metric.

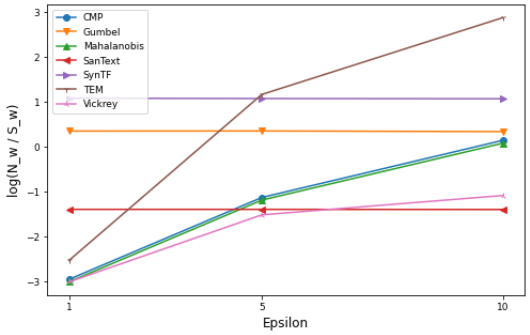
In order to aid in our pursuit to benchmark the privacy-utility trade-off for our selected MLDP methods, we introduce a new metric that aims to capture both the utility of a method and its privacy-preserving capabilities. As such, we utilize the Privacy-Utility Composite (PUC) score, defined as:

$$PUC = \alpha(\frac{100 * Acc}{B-Acc}) + (1-\alpha)\frac{(100-N_w)+S_w+PP+CS+(100-LOW)}{5} \quad (4)$$

Where *(B-)Acc* represents the (baseline) accuracy percentage, and $\{N_w, S_w, PP, CS, LOW\}$ represents the set of privacy metrics we use, as introduced in Section 4.1.5. Scores where lower values are better are subtracted from 100. α is the *privacy-utility tuning parameter*, where one can scale the



(a) IMDb, N_w:S_w ratio



(b) AG News, N_w:S_w ratio

Figure 2: Ratio of N_w to S_w, averaged over three embedding dimensions. Lower ratios correspond to higher plausible deniability. Ratios are shown on the logarithmic scale due to outlier values (i.e., for TEM).

preferred weight placed upon utility and privacy metrics. For example, choosing an α of 0.75 would strongly emphasize the weight of the utility outcomes, whereas the composite privacy score would receive a weight of only 0.25. More generally, we can define the PUC score as follows:

Definition 5.1 (Privacy-Utility Composite (PUC ↑) Score)**.** For a set of (1) utility metrics $\mathcal{U}$, representing percentages compared to a baseline, and (2) a set of privacy metrics $\mathcal{P}$, which can be broken down into metrics $\mathcal{P}\uparrow$ and $\mathcal{P}\downarrow$, and a privacy-utility tuning parameter α, and a max score M:

$$PUC = \frac{\alpha}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} u_i + \frac{1-\alpha}{|\mathcal{P}|}\left(\sum_{p\in\mathcal{P}\uparrow} p_i + \sum_{\tilde{p}\in\mathcal{P}\downarrow}(M-\tilde{p}_i)\right) \quad (5)$$

Note that Equation 5 assumes that all scores are on the same scale, i.e., in the range of $[0, M]$. Scores not on the same scale can be scaled or normalized accordingly. The PUC score also assumes an equal weighting within a metric set, e.g., all utility scores are weighted equally.

It should also be noted that the choice if α is envisioned to be performed *a priori*, or rather, before the design or evaluation of MLDP mechanisms and not as a tunable parameter. In this way, the requirement of privacy protection versus utility preservation should be decided upon beforehand, so that such a balance will be reflected in the analysis of the PUC scoring results.

In Figure 3, we illustrate average PUC scores with three α values. These values are meant to simulate a preference for utility (Figures 3a, 3d), a balanced preference (Figures 3b, 3e), and a preference for privacy preservation (Figures 3c, 3f).

Finally, to explore the relevance of our *CS* privacy metric, we perform a Multiple Linear Regression (MLR) test. We use epsilon (ϵ), N_w, S_w, and *PP* as our predictor variables and *CS* as our response variable. The resulting model and summary of the regression test are shown in Table 4. Most importantly, one can see strong correlations between *CS* and all other variables, notably a strong positive correlation with ε and a strong negative correlation with *PP*. This demonstrates that the choice of ε is a strong indicator of the expected degree of utility in the output privatized text.

$R^2 = 0.905$	coef.	std err	t	P>|t|
const.	126.6834	4.963	25.528	0.000
epsilon (ε)	0.6732	0.192	3.503	0.001
N_w	-0.3158	0.054	-5.818	0.000
S_W	-0.2157	0.028	-7.811	0.000
PP	-0.7728	0.038	-20.337	0.000

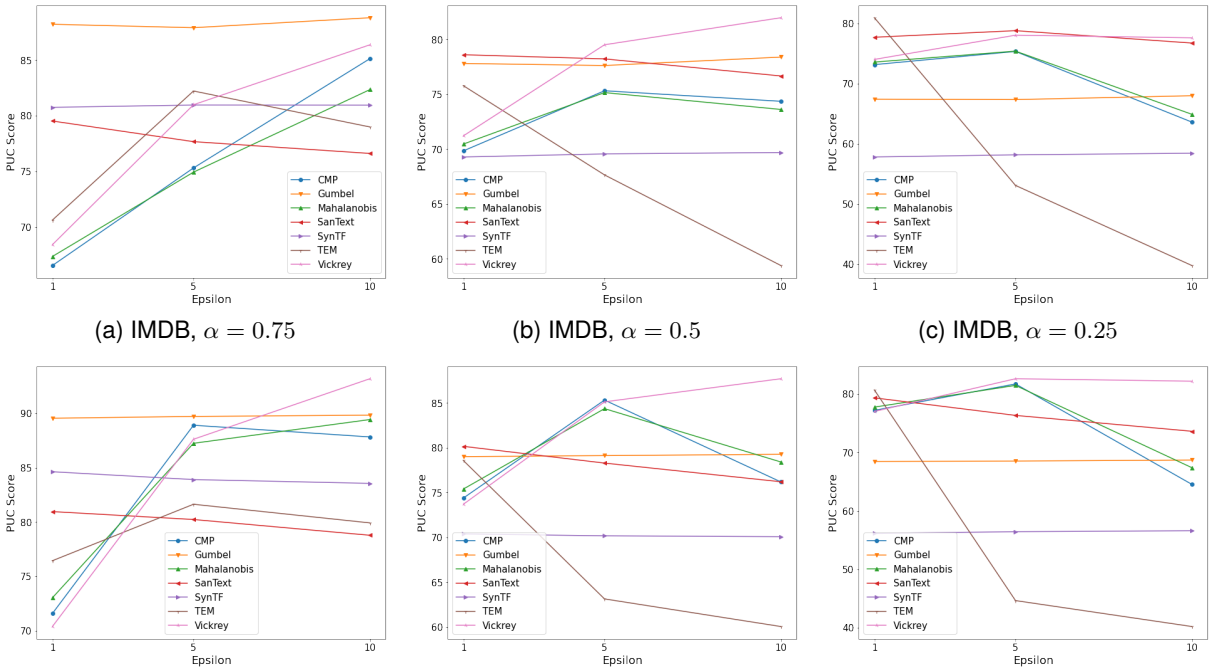
Table 4: MLR to predict the *CS* metric. In general, R^2 measures the goodness of the fit, ranging from 0 to 1. All predictors are statistically significant.

6. Discussion

Effect of DP on Utility While it is reasonable that a lower ϵ value will generally lead to lower utility, a thorough study of our results reveals additional insights. Firstly, some methods prove to be *utility loss invariant* w.r.t. the choice of ϵ. With SynTF, this is explainable by its mechanism design. However, in the case of Gumbel and SanText, their strength in preserving utility across ϵ values is made clear.

We also observe the effect of embedding dimension. Looking at Figure 1, one can see that as embedding dimension increases, Mahalanobis, CMP, and Vickrey all experience drops in accuracy, given a fixed ϵ. It is presumed that this is an artifact of the mechanism design, where utility may begin to deteriorate as more dimensions of noise are added.

Notably, the use of an MLDP mechanism in some cases actually contributes to an increase in accuracy against the baseline, particularly in the case of Gumbel, but also observed in TEM, CMP, and Mahalanobis. Such a phenomenon was observed 13 times in the Topic Classification task and four times in Sentiment Analysis. This finding opens the discussion of MLDP as a "robustness mechanism".



(a) IMDB, $\alpha = 0.75$ (b) IMDB, $\alpha = 0.5$ (c) IMDB, $\alpha = 0.25$

(d) AG News, $\alpha = 0.75$ (e) AG News, $\alpha = 0.5$ (f) AG News, $\alpha = 0.25$

Figure 3: Privacy-Utility Composite (PUC) scores per task, with varying α. The PUC scores were averaged across embedding dimension, and these averages are shown for each epsilon (ϵ) value. The left column with $\alpha = 0.75$ favors utility, the middle with $\alpha = 0.5$ is balanced, and the right with $\alpha = 0.25$ favors privacy.

Privacy Analysis Moving to the analysis of our privacy metrics, we begin with our PD statistics, namely the ratio of N_w to S_w. Much like Figure 1, one can observe in Figure 2 that three methods maintain a near-constant average ratio across ϵ values. With SynTF and SanText, this phenomenon can be attributed to the mechanism design, as they operate differently from the five selected MLDP mechanisms. Furthermore, we see that the ratios of all methods generally follow the trend exhibited by the utility lines in Figure 1, leading us to believe that utility and the $N_w : S_w$ ratio are closely correlated, as best exhibited by the Gumbel mechanism.

An important discussion comes with the clear relationship between a characterization of "effective" perturbation, where one may base the effectiveness of a mechanism in preserving privacy by the level of *plausible deniability* that it provides. Thus, a mechanism that has a high probability of perturbing words (low N_w) and a high *diversity* of output words (high S_w) would present the most attractive option. However, one can observe that the mechanisms with the lowest ratio scores, e.g., the Vickrey mechanism, in Figure 2 also demonstrate lower utility scores in Figure 1. Mechanisms that operate in more of a "balanced" fashion, e.g., the Gumbel mechanism, suffer less from utility drops. This illustrates an important finding regarding mechanism design, namely that plausible deniability and utility preservation must be considered in parallel.

In a similar vein, the MLR analysis gives interesting insights into the connection between privacy and utility. If we assume that N_w and S_w, under a certain ϵ constraint, supplemented with the empirical observation of *PP*, can characterize a general MLDP mechanism well, then one can very well predict CS given any mechanism. This provides a useful link to mechanism design and the (predicted) effect on the preservation of semantics, and ultimately, the effect on utility.

Privacy and Utility in the Same Light Key to this discussion is also an analysis of the composite quantification of privacy and utility, which was aided by our introduced metric: the PUC score. In Figure 3, we see that the PUC score can vary quite significantly with the choice of α. As an example, with an α of 0.75, one can clearly see that the Gumbel Mechanism presents an attractive choice of method, and this is supported by the observed accuracy score of Figure 1. However, if we tune the parameter more towards privacy, this mechanism falls significantly from the top position. Indeed, looking to the privacy results of Gumbel, the observed scores are not on par with the other selected methods. Most notably, S_w tends to be quite low. Similar analyses using the PUC score can be performed, with the goal of tailoring the benchmark interpretation to the privacy preferences of the user.

An immediate challenge with the quantification of the privacy-utility trade-off in a comparative man-

Original:			Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep. In this case I fell asleep 10 minutes from the end, really, really bored and not caring at all about what happened next
Preproc:			sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
Mechanism	d	ϵ	Sentence
CMP	300	1	gr gft expectable chakra grandparent gored magritte noo sniper breakfast meh substantive paternal verifiably viking flute erm striker muddies shane
CMP	300	5	relay neighborhood crime dubai hiroshima vampyre sandal kilcher estimate evers studiously mib cowgirl puzzle coldest stv danube berkley pulitzer del
CMP	300	10	cardiff interpretation efficiency hollywood movie stooped reacting sleep case export asleep minute dillon unable lot bored interfere trainor depressive gunter
SynTF	300	1	meritless devote rat ease_up flick walkway fall departed character diminish asleep mo cease actually in_truth blase not deal materialize future
SynTF	300	5	disconsolate afford rat devote movie base_on_balls downfall gone vitrine fell departed narrow terminal in_truth truly tire non lovingness take_place future
SynTF	300	10	bad cave_in blackleg ease_up movie paseo hang at_peace font precipitate asleep mo stop truly actually bore not manage happen succeeding
TEM	300	1	peckenpahs urchin iñárritu lansburys streaming clout goosebump kissed welcomed maggies iwhippedi waswell occuped damme unbielevable calligraphy cameraman nula sharie british
TEM	300	5	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
TEM	300	10	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
Mahalanobis	300	1	perspicacious peace gf pellet gomer gargan raspberry kursk no prime wisconsin pickier reddin salvific designer clunkers incursion martyr hurd umm
Mahalanobis	300	5	sincere ha hrithik goalkeeper batman innes pole astral bellucci visa disfigured clan wale geometry faceoff simon sharia humperdink von faulty
Mahalanobis	300	10	sad summary rating age movie mum monarch asleep psychiatric fell asleep goalkeeper talker silver prototype improvising office caring thigpen declined
SanText	300	1	sorry gave rating give granddaughter walk fall asleep case fell asleep minute end really workhorse generate agreeing caring happened next
SanText	300	5	sorry gave rating give movie walk fall asleep hypothetically huggable selina memorably end crucially really starstruck insisting caring happened next
SanText	300	10	sorry gave rating give movie begun fall asleep concluded fell asleep minute end really really bored be caring happened next
Gumbel	300	1	embarrassed give indicating gave filmed anyway though awake reason dropped woke minute end definitely certainly tired though evera happen week
Gumbel	300	5	disappointed put rating giving movie walking fall awake example falling asleep equalizer ended really know bored although elderly happened expected
Gumbel	300	10	ashamed giving rating given starred walked coming sleep example slid fortunately came however obviously certainly bored be loving happening take
Vickrey	300	1	dah mayoral herein wachowski address ee corneau blazing ketchup observatory curled verdi thematic zen materialises ishwar wrestlemania nicholsons sonja interference
Vickrey	300	5	pepe pota eavesdrops hatching stunt yeop traumatizing takoma detained factly picher hitch light englund encyclopedia glanced calcium ditzy pasta chromosome
Vickrey	300	10	miserable assertion plunging invocation jerker sabre competing appetizer homicide dated suspended sanchez levy go consistency scene entertained flawed dreamt cbs

Table 5: Example text output on the IMDb dataset, for all evaluated mechanisms and ε values on 300-dimensional GloVe embeddings. The original dataset text, as well as the preprocessed text, is given.

ner traces back to the foundations of Metric Differential Privacy. With the generalization of DP to metrics, the comparability of the ε parameter across mechanisms becomes more difficult, especially for those operating in different metric spaces. Although we evaluate our selected MLDP mechanisms on discrete ε values, a more calibrated evaluation presents a concrete opportunity for improving the comparability of word-level MLDP mechanisms.

The Question of Metrics Our work highlights the need for a qualified and agreed upon set of metrics, which are necessary in order to evaluate word-level MLDP metrics on a uniform platform. In addition, this need for evaluation extends beyond the word-level to all DP NLP methods. The challenges this brings are numerous, rooted in the core challenge discussed above, i.e., the comparability of ε. In this work, we aim to begin the discussion with a base set of metrics, which provide the foundation for further metrics, as well as the opportunity to validate the efficacy of these metrics.

To start such validation, we critically view some of the privacy metrics proposed here. The privacy metric of *LOW* presents an interesting point of inquiry, as this score varies quite significantly between mechanisms and experiment runs. No discernible interpretation, therefore, can be drawn; thus, an analysis of the usefulness of such a score is a topic of future investigation; therefore, further studies into the usefulness of this metric, as well as other lexical- or syntactic-based metrics, would be well-served. In addition, it becomes very important for the field of evaluating text privatization to agree upon a standard set of privacy metrics, something that currently does not exist. Such standardization would be paramount in unifying the validation and evaluation of privacy in NLP.

Another dimension of evaluation and metrics not directly covered in this work is that of *semantic coherence* and *readability*. Although our CS and PP metrics capture to a degree the "closeness" of the perturbed text to the original, a closer look at the perturbed outputs (see Table 5) illustrates that there is still much room for improvement. Therefore, going forward in evaluating DP mechanisms, a greater emphasis should be placed on producing readable, coherent privatized outputs. This is also supported by the analysis of Mattern et al. regarding the question of optimal text privatization.

7. Conclusion

In this work, we conduct a comparative analysis of seven word-level Differential Privacy mechanisms, motivated by a lack of uniformity in the evaluation of word-level MLDP methods. We design a multi-dimensional experimental setup, which evaluates our chosen methods on two NLP tasks, with three ϵ parameters and three GloVe embedding dimensions, resulting in a total of 126 data points. To aid in the analysis, we employ a combination of utility and privacy metrics, as well as a novel composite score to quantify the interplay between the two. Finally, we include a discussion and analysis of the observed results, with the goal of providing a basis upon which future works on word-level MLDP evaluation for NLP can build. To this end, we provide a full replication repository, which can be found at `https://github.com/sjmeis/MLDP`.

As additional points of future work, we suggest (1) a more in-depth critique of the merits of word-level MLDP, (2) a focus on improving syntactic and semantic coherence in DP text perturbations (see Table 5), (3) the evaluation of MLDP mechanisms on a more diverse set of tasks and model architectures, and (4) further refinements on the metric-driven quantification and understanding of the privacy-utility trade-off in the NLP domain.

The results of our comparative analysis show that the classic argument of "higher privacy, lower utility, and vice versa" is considerably more complex, particularly in the realm of text. The evaluation of word-level MLDP methods is a start to tackling this question, and its implications provide the impetus to the continued study of privacy in NLP.

8. Acknowledgements

The authors would like to thank Maulik Chevli for his valuable contributions to this work.

9. Limitations and Ethics Statement

While the aim of our work was to provide a uniform and fair evaluation for word-level MLDP methods, a clear limitation comes with the selection of our experiment parameters, or rather, those that were left out. Firstly, only pre-trained GloVe embeddings were utilized; other models such as Word2Vec or FastText were not included. In addition, the choice of ϵ can be perceived as limiting, especially the lack of uniformity in ϵ selection, and the resulting effects. This is especially crucial with MLDP, as the underlying distance metric for each mechanism affects the scale of noise added. Finally, the limitation of computational resources did not allow us to fine-tune algorithm-specific parameters; thus, we chose a single value for such parameters.

Other limitations include our choice of privacy metrics. Empirical Privacy (degradation of adversarial performance) was not measured. Furthermore, our chosen metrics included previously unused metrics, such as *LOW*. While these metrics were seen to be useful for illustrative purposes, validation of them as useful metrics would be an excellent point of future work.

On the note of metrics, our proposed PUC Score assumes that all individual metrics included in the composite scoring are weighted equally, an assumption that may or may not reflect the preferences or requirements of real-world practitioners. Thus, further work in the refinement of the composite metric to allow for such flexibility is needed.

As a benchmarking study, a final limitation comes with the fact that we did not benchmark time or resource consumption for each of our experimental runs. A quantification of the time needed, as well as the computational overhead, to perform word-level MLDP would be very useful in completing the picture we present in this work.

Regarding ethical implications, the core of this study looks to an increasingly important topic of societal relevance, namely that of data privacy. In this light, we hope that our work contributes to the principle of respecting privacy in the processing of text data, particularly that which may contain sensitive information.

An ethical consideration to note is using pre-trained word embedding models as the basis for word-level MLDP. Possible ethical concerns with such models have been pointed out (Bolukbasi et al., 2016; Papakyriakopoulos et al., 2020; Manzini et al., 2019), and the effect of bias in these models was not tested by our study.

10. Bibliographical References

Mário Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. 2018. Local differential privacy on metric spaces: optimizing the trade-off with utility. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 262–267. IEEE.

Gilles Barthe, Marco Gaboardi, Justin Hsu, and Benjamin Pierce. 2016. Programming language techniques for differential privacy. *ACM SIGLOG News*, 3(1):34–53.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

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

Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: high utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 883–890. SIAM.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13*, pages 82–102. Springer.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy

for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*, pages 123–148. Springer International Publishing.

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*, pages 178–186.

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)*, pages 210–219.

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, Online. Association for Computational Linguistics.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2023. Differentially private natural language models: Recent advances and future directions. *arXiv preprint arXiv:2301.09112*.

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.

Yilin Kang, Yong Liu, Ben Niu, Xinyi Tong, Likun Zhang, and Weiping Wang. 2020. Input perturbation: A new paradigm between central and local differential privacy. *arXiv preprint arXiv:2002.08570*.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

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.

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.

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 '20, page 1813–1816, New York, NY, USA. Association for Computing Machinery.

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.

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, and Aurélien Bellet. 2022. Fair NLP models with differentially private text encoders. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6913–6930, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, pages 615–621.

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.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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)*, pages 3–18. IEEE.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

Jingye Tang, Tianqing Zhu, Ping Xiong, Yu Wang, and Wei Ren. 2020. Privacy and utility tradeoff for textual analysis via calibrated multivariate perturbations. In *Network and System Security*, pages 342–353, Cham. Springer International Publishing.

Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, pages 273–281.

Benjamin Weggenmann and Florian Kerschbaum. 2018. SynTF: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 305–314.

Nan Xu, private release of text Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021a. Density-aware differentially private textual perturbations using truncated gumbel noise. *The International FLAIRS Conference Proceedings*, 34.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021b. On a utilitarian approach to privacy preserving text generation. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 11–20, Online. Association for Computational Linguistics.

Zekun Xu, Abhinav Aggarwal, private release Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using regularized mahalanobis metric. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 7–17.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.