The Limits of Word Level Differential Privacy

Justus Mattern
RWTH Aachen University
justus.mattern@rwth-aachen.de

Benjamin Weggenmann
SAP Security Research
benjamin.weggenmann@sap.com

Florian Kerschbaum
University of Waterloo
florian.kerschbaum@uwaterloo.ca

Abstract

As the issues of privacy and trust are receiving increasing attention within the research community, various attempts have been made to anonymize textual data. A significant subset of these approaches incorporate differentially private mechanisms to perturb word embeddings, thus replacing individual words in a sentence. While these methods represent very important contributions, have various advantages over other techniques and do show anonymization capabilities, they have several shortcomings. In this paper, we investigate these weaknesses and demonstrate significant mathematical constraints diminishing the theoretical privacy guarantee as well as major practical shortcomings with regard to the protection against deanonymization attacks, the preservation of content of the original sentences as well as the quality of the language output. Finally, we propose a new method for text anonymization based on transformer based language models fine-tuned for paraphrasing that circumvents most of the identified weaknesses and also offers a formal privacy guarantee. We evaluate the performance of our method via thorough experimentation and demonstrate superior performance over the discussed mechanisms.

1 Introduction

Computational authorship attribution approaches ranging from rule-based methods measuring character-level n-gram frequencies (Kešelj et al., 2003) to models incorporating deep learning (Shrestha et al., 2017) make it possible to identify the authors of a given text. While these technologies enable valuable applications such as supporting historians in their research, they can potentially be exploited by attackers to identify the originators of sensitive data and thus diminish the privacy of individuals. To protect the anonymity of users whose data is being shared online and used by companies and researchers, methods that anonymize the writer of given texts are necessary and of interest within the research community and a variety of industries, specifically those handling personal information such as healthcare or financial services.

Previous work in the field of authorship obfuscation mainly focuses on two different tasks, namely learning anonymous textual vector representations for downstream tasks (Coavoux et al., 2018a; Weggenmann and Kerschbaum, 2018; Fernandes et al., 2019; Mosallanezhad et al., 2019; Beigi et al., 2019) and the development of mechanisms that transform the input sentence to remove properties revealing the author and thus output human-readable text. Works within the second category (Feyisetan et al., 2019, 2020; Xu et al., 2020b; Bo et al., 2021) typically follow a common *word level* framework which is characterized by the differentially private individual perturbation of word embeddings and the subsequent sampling of new words that are close to the perturbed vectors in the embedding space. Also, the majority of recent work proposing new methods for authorship obfuscation deals with the optimization and calibration of noise sampling mechanisms (Xu et al., 2020a) or the definition of new distributions to sample noise from (Feyisetan et al., 2019) as opposed to the development of entirely new methods.

In this paper, we thoroughly investigate the capabilities of word level anonymization from the theoretical perspective of differential privacy (DP) (Section 3.1), in terms of the language quality of its output (Section 3.2) as well as from a utilitarian perspective considering the ability to protect the privacy of people whose data is being used. Specifically, we extend the experimentation in papers proposing the discussed methods by testing their capability to mitigate deanonymization attacks using state-of-the-art methods on the widely used IMDb movie review and Yelp business review datasets (Section 5). We find that the technical constraints applied to fulfill DP in the local model cause strong limitations, and, more importantly, observe that,

Findings of the Association for Computational Linguistics: NAACL 2022, pages 867 - 881

despite the formal guarantees, such methods offer little protection against advanced deanonymization attacks. For this reason, we advocate for approaches granting more flexibility to the text generation process and, motivated by experiments showing that human rewritings of texts gathered through crowdsourcing successfully anonymize the original authors (Almishari et al., 2014), propose an anonymization approach based on paraphrasing (Section 4) that maintains the advantages and the theoretical privacy guarantee of the discussed methods, evades most of the identified drawbacks and outperforms word level mechanisms in our experiments.

2 Background

The majority of proposed text anonymization methods rely on a common framework that applies DP on a per-word level by perturbing individual word embeddings (Feyisetan et al., 2019, 2020; Xu et al., 2020b,a, 2021). In this section, we introduce the concept of DP and give an overview of the commonly used word level framework.

2.1 Differential Privacy

DP has been introduced by Dwork et al. (2006) under the name ϵ-*indistinguishability*. Its goal is to give semantic privacy by quantifying the risk of an individual that results from participation in data collection. In the original, *central model*, we assume the collected data is stored in a central database with one record per participant. If we consider *adjacent* databases that differ by at most one record (pertaining to one individual), a differentially private query on both databases should yield matching results with similar probabilities, i.e., answers that are probabilistically *indistinguishable*. This is achieved via *random mechanisms* on the universe of datasets $\mathcal{D}$ that return noisy query results, thus masking the impact of each individual.

Definition 1 (ϵ-DP) *Let $\epsilon > 0$ be a privacy parameter. A random mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ fulfills* ϵ-DP *if for all adjacent databases $D, D' \in \mathcal{D}$, and all sets of possible outputs $R \subset \operatorname{supp} \mathcal{M}$,*

$$\Pr[\mathcal{M}(D) \in R] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in R].$$

To make a query function $f : \mathcal{D} \to \mathcal{R}$ differentially private, noise is calibrated to the query's *sensitivity*, i.e. its maximal change over all pairs of adjacent datasets $D \sim D' \in \mathcal{D}$. For instance, the L2-sensitivity as used for the Planar Laplace mechanism (Chatzikokolakis et al., 2013; Andrés et al., 2013; Koufogiannis et al., 2015) is

$$\Delta_2 f := \max_{D \sim D'} \|f(D) - f(D')\|_2.$$

In the *local model* (Duchi et al., 2013), noise is added locally at the data source, before the data is collected and stored in a central database. A basic example is randomized response (Warner, 1965), where each survey participant either provides a truthful or a random answer depending on the flip of an (unbiased) coin. The local model makes the strong assumption that any two inputs are considered adjacent, which often makes it difficult to achieve a satisfying privacy-utility trade-off.

2.1.1 Generalization with metrics

A limitation with DP is that the indistinguishability is achieved between two inputs on a per-record level regardless of their actual values. This can be especially problematic in the local model, where each user might just submit one single record, in which case a DP mechanism with small privacy parameter ϵ would enforce each submitted record to be indistinguishable from any other, thus rendering the collected data essentially useless. Chatzikokolakis et al. (2013) argue that in some scenarios, the (in)distinguishability between two databases as enforced by a privacy mechanism should depend on the values themselves instead of the number of differing records. They hence propose a generalized notion of *privacy on metric spaces* where a mechanism run on *nearby* elements results in *similar* output probabilities:

Definition 2 (Metric privacy) *Let $\epsilon > 0$ be a privacy parameter. On a metric space $(\mathcal{X}, d)$, a mechanism $\mathcal{M}$ satisfies* ϵd-privacy *if for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and all $R \subset \operatorname{supp} \mathcal{M}$,*

$$\Pr[\mathcal{M}(\boldsymbol{x}) \in R] \leq e^{\epsilon \cdot d(\boldsymbol{x}, \boldsymbol{x}')} \cdot \Pr[\mathcal{M}(\boldsymbol{x}') \in R].$$

In other words, the indistinguishability level of two points $\boldsymbol{x}, \boldsymbol{x}'$ is bounded by ϵ times their distance.

Note that we recover the original notion of central DP on the space of databases $\mathcal{X} = \mathcal{D}$ if we use the *record-level edit distance* $d_{\pm 1}$, as datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ differ by at most one record if and only if $d_{\pm 1}(\boldsymbol{x}, \boldsymbol{x}') \leq 1$. Similarly, the local model is obtained for $d(\boldsymbol{x}, \boldsymbol{x}') \equiv 1$. This motivates the following broader and formal definition of adjacency:

Definition 3 *In a metric space $(\mathcal{X}, d)$, we call two inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$* adjacent (with respect to d) *if $d(\boldsymbol{x}, \boldsymbol{x}') \leq 1$. We write this as $\boldsymbol{x} \sim_d \boldsymbol{x}'$ (or $\boldsymbol{x} \sim \boldsymbol{x}'$ if d is understood from the context).*

2.2 Word perturbations for anonymization

The methods investigated in this paper apply word embedding perturbation mechanisms to change individual words in a sentence, following ϵd-privacy with a distance metric defined for sentences $\boldsymbol{x}, \boldsymbol{x}'$. In essence, the common word level framework works as follows: Given an input sentence $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, each token x_i is mapped to an n-dimensional pretrained word embedding $\phi(x_i)$. Subsequently, an n-dimensional noise vector η is sampled from a multivariate probability distribution $p_\epsilon(\eta)$ and added to the word embedding to obtain a noisy vector $\hat{\phi}_i$. The current word x_i then gets replaced by a word x_i' whose embedding $\phi(x_i')$ is close to the noisy embedding $\hat{\phi}_i$. Given a distance metric d, commonly $d(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{n} \|\phi(x_i) - \phi(x_i')\|$ for sentences $\boldsymbol{x}, \boldsymbol{x}'$ of the same length, the mechanism fulfills ϵd-privacy. The general mechanism is outlined in Algorithm 1 and the proofs are outlined in the referenced papers.

Algorithm 1: Word level DP framework

Input : Text $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, parameter ϵ
Output: Anonymized text $\boldsymbol{x}' = (x_1', x_2', ..., x_n')$
for $i \in \{1, 2, \ldots, n\}$ **do**
  Compute embedding $\phi_i = \phi(x_i)$
  Sample noise $\eta \sim p_\epsilon(\eta)$
  Compute perturbed embedding $\hat{\phi}_i = \phi_i + \eta$
  Find near word x_i' within embedding space
  Insert x_i' for x_i in the output

3 Limitations of word level privacy

DP mechanisms operating on a word-by-word basis follow a comparably simpler and more straightforward algorithmic approach than deep learning models for text anonymization. This has many advantages such as lower computational expense as well as the mechanism's independence of the target dataset and domain: Most deep learning based approaches need to be trained for each dataset and set of authors individually as they require author labels to construct adversarial training objectives (Shetty et al., 2018; Xu et al., 2019). In contrast, the approaches discussed in this paper are dataset-independent and can thus be deployed immediately without a need for further training for new authors and datasets.

The simple methodology does however have its shortcomings as well. In this section, we examine these weaknesses from a theoretical standpoint taking into account both DP properties and properties of the language output before assessing their effects experimentally in Section 5.

3.1 DP related constraints

We consider a mechanism $\mathcal{M}$ that operates on a text $\boldsymbol{x} = (x_1, \ldots, x_n)$ on a word-by-word basis, i.e., $\mathcal{M}(\boldsymbol{x}) = (\mathcal{M}(x_1), \ldots, \mathcal{M}(x_n))$.

Length constraints A word level mechanism $\mathcal{M}$ will produce an output that has the same length as its input. However, typical texts and sentences come in varying lengths, say $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}' = (x_1', \ldots, x_m')$ with $n \neq m$. Now if we consider an outcome set Z_n consisting of all length-n sequences (including $\boldsymbol{x}$), we obtain

$$1 = \Pr[\mathcal{M}(x) \in Z_n] \not\leq e^{\epsilon} \Pr[\mathcal{M}(x') \in Z_n] = 0.$$

This contradicts the definition of pure DP and in case of approximate DP (cf. Definition 1) would require $\delta = 1$ which is clearly not negligible.

To comply with these strong DP requirements, word level DP mechanisms such as Feyisetan et al. (2019, 2020) commonly simply limit the privacy guarantee to cover only sequences Z_n of a fixed length n, i.e., no formal guarantee among sentences of different lengths is provided. Consequently, the output is also fixed to length n, which affects the language capabilities of such mechanisms and severely limits the scope and expressiveness of the resulting sentences, particularly for human readers.

Linear growth of privacy budget For an ϵ-DP mechanism $\mathcal{M}$, its output probabilities given two adjacent inputs have to be bounded by $\exp(\epsilon)$. Suppose $\mathcal{M}$ processes each word x_i of a text $\boldsymbol{x} = (x_1, \ldots, x_n)$ independently, using a fixed-length output strategy as described in the preceding section, with a given output $\boldsymbol{z} = (z_1, \ldots, z_n)$. Then $\Pr[\mathcal{M}(\boldsymbol{x}) = \boldsymbol{z}] = \prod_{i=1}^{n} p_i$ where $p_i := \Pr[\mathcal{M}(x_i) = z_i]$. Similarly, a second text $\boldsymbol{x}'$ has output probabilities $p_i' = \Pr[\mathcal{M}(x_i) = z_i]$, so we have $p_i \leq e^{\epsilon} p_i'$, and hence

$$\begin{aligned}\Pr[\mathcal{M}(\boldsymbol{x}) = \boldsymbol{z}] &= \prod_{i=1}^{n} p_i \leq \prod_{i=1}^{n} e^{\epsilon} p_i' \\ &= e^{n\epsilon} \Pr[\mathcal{M}(\boldsymbol{x}') = \boldsymbol{z}].\end{aligned}$$

Therefore, the total privacy budget required by $\mathcal{M}$ to privatize the entire sequence is bounded by $n\epsilon$ and thus may grow linearly with its length.

Metric privacy *hides* this effect *in* the metric, since deviations in the mechanism's output proba-

bilities are bounded by $\exp(\epsilon d(\boldsymbol{x}, \boldsymbol{x'}))$. By choosing a metric d that grows larger as the length of sentences increases, strong deviations can now be captured by the metric d, so the privacy budget ϵ as its co-factor *appears* smaller. For instance, Feyisetan et al. (2020) use a metric $d(\boldsymbol{x}, \boldsymbol{x'}) = \sum\|\phi(x_i) - \phi(x'_i)\|$ for strings based on embeddings ϕ, which results in more summands and thus larger distances for longer strings, but not necessarily larger distances for different writing styles: Consider the following sentence pairs $(\boldsymbol{x}, \boldsymbol{x}')$ and $(\boldsymbol{y}, \boldsymbol{y}')$ written by two authors each:

$\boldsymbol{x}$ = "Today I feel great"
$\boldsymbol{x}'$ = "I feel great today"
$\boldsymbol{y}$ = "Today I feel great and will get a coffee"
$\boldsymbol{y}'$ = "I feel great and will get a coffee today"

Given a non-degenerate metric d, we have both $d(\boldsymbol{x}, \boldsymbol{x}'), d(\boldsymbol{y}, \boldsymbol{y}') > 0$ since the sentences are syntactically different. One could infer that the author of the first sentence within both pairs tends to put expressions of time in the beginning whereas the other author places them at the end, but beyond that, there are arguably no differences in terms of writing style or author-revealing information one could deduce from both sentence pairs. Yet, we will likely have $d(\boldsymbol{x}, \boldsymbol{x}') < d(\boldsymbol{y}, \boldsymbol{y}')$ due to the induced growth of the distance for longer sentences. Hence, while the distance metric does reflect differences between sentences in a somewhat meaningful way, it is prone to absorb the actual privacy loss even if the writing style is almost unchanged, thus leading to values of ϵ that are *perceived* as small.

Shortcomings of the local model In many likely scenarios for authorship obfuscation methods, the intention is to share obfuscated texts with other, benign entities for further processing. For a DP mechanism, this essentially corresponds to the local model where it transforms each text individually to an obfuscated output. The assumption then is that the obfuscation allows only privacy-insensitive processing so that subsequent results and inferences do not harm the privacy of the texts' authors.

Note that the DP guarantee in the local model differs substantially from what is expressed by the definition in the original central model: A central DP mechanism would aggregate the texts (records) from multiple individuals into a single result. By the definition of adjacency, central DP hides the impact of each individual's contribution in the result by making it probabilistically indistinguishable (as determined by ϵ) whether an outcome was obtained with or without an individual's data. In contrast, for local DP, any two inputs are considered adjacent, so by definition, it needs to be indistinguishable whether an output was produced by one input or another. This strong condition makes it thus questionable if such data is still useful for an analyst.

Due to the nature of local DP, it typically introduces large amounts of noise and requires large amounts of data to still get meaningful results (Wood et al., 2020). A workaround often used in practice when only limited data is available is to use a larger privacy budget ϵ than one would normally consider sufficiently privacy-preserving in the central model (Qin et al., 2016; Desfontaines, 2021). While this does permit the obfuscated data to remain useful to a benign analyst, it may also be useful to an attacker to infer privacy-sensitive information, as the formal guarantee of local DP does not specifically prevent such undesired or malicious inferences, especially when ϵ is large.

To alleviate the strictness and implications imposed by the local model, some approaches refer to metric privacy (Chatzikokolakis et al., 2013) as generalization of the original definition. Metric privacy (cf. Definition 2) brings about a change in the definition how the privacy loss ϵ is interpreted in relation to the introduced metric and normally leads to seemingly smaller ϵ values; however, changing to metric privacy by itself does not imply any change to the inner workings of the mechanism. We hence argue that it is less an improvement, but more a relaxation of the privacy guarantee that still shares the same fundamental criticism of local DP, e.g., our observation at the start of this section where the metric grows with the length of the text and thus hides the linear growth of the privacy budget.

3.2 Language constraints

Aside from weaknesses concerning the privacy guarantee of DP, mechanisms operating on a per-word level pose two significant shortcomings in terms of their language generation capability. First, smaller privacy budgets resulting in stronger noise added to the original data tend to cause a high amount of grammatical errors. Secondly, the lack of syntactic changes to the original sentences caused by the nature of such mechanisms considerably limits the linguistic variety and thus opportunities to deceive an adversary and provide anonymity

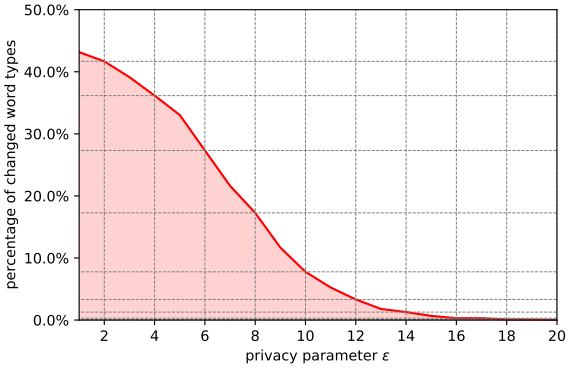

Figure 1: Percentage of word type changes caused by the mechanism introduced by Feyisetan et al. (2020)

for the authors of the texts.

Grammatical errors increase as privacy budget shrinks Word level mechanisms perturb every token x_i in a sentence independently of the rest of the text as opposed to common autoregressive sequence-to-sequence models where $p(x_i) = p(x_i \mid x_{i-1}, \ldots, x_1)$. This makes it difficult to maintain consistency and renders them unable to rectify grammatical errors induced by replacing a word with one of a different word type, e.g., a noun with an adjective. To estimate the effect of this, we approximate the likelihood of word type exchanges for various ϵ values: Using the WordNet database[1] (Miller, 1995; Fellbaum, 2010), we assign words from the GloVe vocabulary (Pennington et al., 2014) one or multiple of the word type labels *adjective*, *adverb*, *noun* and *verb*. Subsequently, we apply the word perturbation mechanism proposed by Feyisetan et al. (2020) on a randomly selected set of 1,000 tokens and use the assigned type labels to measure whether the word type was changed[2] or not.

As Fig. 1 shows, a significant percentage of word type changes occur even when using comparably large ϵ values such as 8 or 10 that grant only little privacy protection according to our evaluation in Section 5: With 17.3% and 7.8% of word types being changed with the respective epsilon values, a word type change and thus most likely a grammatical error would be induced at every 5.8th and 12.8th token, respectively.

[1]Terms of use and license information: Appendix A.1

[2]In case of multiple word type labels for a single token (e.g. noun and verb for "escape"), we only interpreted the perturbation as a word type change if the sets of word types of the original word and the new word were disjoint.

Lack of syntactic changes As described in Section 3.1, operating on a word-by-word basis causes severe limitations to the format of the perturbed output sentences. Due to the imposed inflexibility of the text generation process, the discussed mechanisms lack the ability to rewrite given sentences by changing their syntactic properties such as word positioning and sentence length and thus mostly have to rely on lexical changes for obfuscating author-revealing features, which is highly unfavorable. For instance, if a person's writing style is characterized by heavy use of subordinate clauses resulting in very long sentences, it may be more effective to shorten sentences than merely changing individual words.

Due to these limitations, word level methods may never achieve proper anonymization, as even syntactic features alone without any semantic information are sufficient for authorship identification: Notably, Tschuggnall and Specht (2014) show that, given a collection of syntactic trees of texts written by various authors, individual style profiles can be learned to infer the authors of unseen sentences. Moreover, learned representations of syntax trees have proven to be effective for various authorship attribution tasks (Hitschler et al., 2017; Zhang et al., 2018; Jafariakinabad et al., 2019). Consequently, an effective anonymization mechanism should be able to change the syntactic properties of its input texts in order to take away important clues that adversaries could exploit to identify authors.

4 Anonymization through paraphrasing

While existing works on text anonymization that focus on word level perturbations represent very important contributions, they have some significant weaknesses as described in Section 3. In the following, we attempt to address the identified problems by proposing fine-tuning of large language models for paraphrasing as an alternative text anonymization method.

4.1 Generating paraphrases

Authorship obfuscation has been framed as a paraphrasing problem in various works with different attempts to generate adequate rewritings (Rao and Rohatgi, 2000; Keswani et al., 2016; Bevendorff et al., 2019; Mahmood et al., 2019). While computational approaches do not always show satisfying results, Almishari et al. (2014) demonstrate that rewritings of reviews gathered through crowdsourc-

ing reflect strongly different stylometric features from the source reviews while preserving the original content and concealing the author successfully.

Crowdsourcing is highly laborious and cannot always be applied in real-world scenarios. Therefore, we aim at imitating the rewriting behavior of humans through a large-scale transformer-based (Vaswani et al., 2017) language model: We fine-tune GPT-2 (Radford et al., 2019) to generate paraphrases following the training procedure introduced by Witteveen and Andrews (2019). The Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) provides training data consisting of pairs of sentences with five crowdsourced labels, each indicating whether the two sentences are semantically entailed or not. We construct a paraphrase dataset by only keeping sentence pairs with all five labels indicating entailment.

4.2 Balancing privacy and utility

As pointed out by Brennan et al. (2012), the black box nature of authorship obfuscation via round-trip and consequently also monolingual translation affects controllability of our system negatively. Therefore, in the following we demonstrate how varying the temperature in the word sampling stage of GPT-2 can be used to inject noise into our model, hereby balancing the privacy-utility trade off.

In an autoregressive generative model, an output text $\boldsymbol{x} = (x_1, \ldots, x_n)$ is generated by sampling the next word x_i from conditional probabilities $\boldsymbol{p}_i = p(x_i \mid x_1, \ldots, x_{i-1}, \boldsymbol{z})$ modeled by the decoder network, where $\boldsymbol{z}$ is context information (e.g., representing an encoding of the input sentence to be obfuscated) to initialize the decoder. The vector $\boldsymbol{p}_i = (p_{i,j})_{j=1}^{|\mathcal{V}|}$ represents the probabilities of producing the j-th word v_j of the predefined vocabulary $\mathcal{V}$ at the i-th position in the sequence. The probabilities are typically obtained through the softmax function from a logit vector $\boldsymbol{u}_i \in \mathbb{R}^{|\mathcal{V}|}$ in the last layer of the decoder, which can be controlled by a temperature parameter $T > 0$ as follows:

$$p_{ij} := \mathrm{softmax}(\boldsymbol{u}_i) = \frac{\exp\left(\frac{u_{i,j}}{T}\right)}{\sum_k \exp\left(\frac{u_{i,k}}{T}\right)} \quad (1)$$

A higher temperature T results in a smoother distribution that brings the resulting probabilities of all words closer together and thus impacts the variability and probabilities of the resulting sentences. In our experiments in Section 5, we vary the temperature when sampling text to evaluate this effect.

Sampling from softmax **as differential privacy mechanism** Note that sampling from the softmax distribution with temperature T can be interpreted as a DP mechanism, namely as an instance of the *Exponential mechanism* by McSherry and Talwar (2007). It applies to both numerical and categorical data and requires a "measure of suitability" for each possible pair of input and output values:

Definition 4 (Quality function) *A map $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called* quality function from $\mathcal{X}$ to $\mathcal{Y}$ *where we interpret the value $q(x, y)$ as measure of suitability of an output $y \in \mathcal{Y}$ for a given input $x \in \mathcal{X}$. The* sensitivity Δ_q *of the quality function q is its largest possible difference given two adjacent inputs, over all possible output values:*

$$\Delta_q := \max_{y \in \mathcal{Y}} \max_{x_1 \sim x_2} \big(q(x_1, y) - q(x_2, y)\big)$$

Given an admissible rating function q with finite sensitivity Δ_q, the Exponential mechanism is defined as follows:

Definition 5 (Exponential mechanism) *Let $\epsilon > 0$ be a privacy parameter, and let $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a rating function. The* Exponential mechanism *is a random mechanism $\mathcal{E} : \mathcal{X} \to \mathcal{Y}$ that is defined by the probability distribution function*

$$\Pr[\mathcal{E}(x) = y] = \frac{\exp\left(\frac{\epsilon}{2\Delta_q} q(x, y)\right)}{\int_{y'} \exp\left(\frac{\epsilon}{2\Delta_q} q(x, y')\right) \mathrm{d}y'}.$$

A discrete version of the Exponential mechanism for countable $\mathcal{Y}$ can be obtained by replacing the integral with a sum; it is thus defined by the probability mass function

$$\Pr[\mathcal{E}(x) = y] = \frac{\exp\left(\frac{\epsilon}{2\Delta_q} q(x, y)\right)}{\sum_{y'} \exp\left(\frac{\epsilon}{2\Delta_q} q(x, y')\right)}. \quad (2)$$

The Exponential mechanism $\mathcal{E}$ fulfills ϵ-DP as shown by McSherry and Talwar (2007, Theorem 6).

By comparing Eqs. (1) and (2), we immediately recognize that sampling from the softmax probabilities $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,|\mathcal{V}|})$ amounts to running an instance of the Exponential mechanism with $\epsilon = 2\Delta_q / T$ and the quality function determined by the logits vector $\boldsymbol{u}_i \in \mathbb{R}^{|\mathcal{V}|}$ as

$$q_i((x_1, \ldots, x_{i-1}, \boldsymbol{z}), v_j) = u_{i,j}, \quad 1 \le j \le |\mathcal{V}|,$$

at each iteration i when sampling the next word x_i. Therefore, our generative paraphrasing model naturally forms a locally differentially private mechanism that also enjoys formal privacy guarantees:

The total privacy budget amounts to $\epsilon \cdot n$ where n is the length of the generated paraphrase. Finally, note that we obtain a finite sensitivity $\Delta_q \leq 1$ by constraining the decoder layer so that the logits in its output fulfill $0 \leq u_{i,j} \leq 1$.

While this approach is still subject to the implications of the local model, and its total privacy budget $\epsilon \cdot n$ may still grow linearly in the length of the produced output, it avoids the language and fixed output length constraints of previous word level privacy mechanisms stated in Section 3.

5 Evaluation

We argue that despite formal guarantees, the privacy preservation capabilities of mechanisms that are deployed in real world applications should also be tested from a practical standpoint. Previous works measure anonymization capabilities using a variety of evaluation metrics: Feyisetan et al. (2020) use the privacy auditor proposed by Song and Shmatikov (2019), whereas Xu et al. (2021) measure the ability of an adversary to reconstruct the original sentence, and Xu et al. (2020b,a) count the amount of changed words.

Unfortunately, these methods are rarely tested under the scenario of a strong attacker aiming to identify the authors of the obfuscated texts. While Feyisetan et al. (2019) measure the identification performance of an authorship attribution model, their adversary (Koppel et al., 2011) only relies on counting character 4-grams and does not adequately reflect the capabilities of a strong attacker who can train more powerful classifiers. Besides, attacks are almost always evaluated only in a static (non-adaptive) setting, meaning that the attack model is only trained on the original data and cannot adapt to the perturbed data. Since any serious method should avoid "security (or privacy) by obscurity", we must assume that the obfuscation mechanism is known to the attacker who can easily create perturbed data themselves.

In the following evaluation, we consider two exemplary methods following the word level framework, namely perturbing Euclidean GloVe embeddings (Pennington et al., 2014) through Laplace noise as proposed by Feyisetan et al. (2020), the perturbation of hierarchical Poincaré embeddings (Nickel and Kiela, 2017) through hyperbolic noise as proposed by (Feyisetan et al., 2019), as well as our paraphrasing approach proposed in Section 4. To address the discussed issues in previous evaluation methodologies, we employ recent state-of-the-art methods to compare the privacy-utility trade-offs and analyze the performance of the approaches not only in a static, but also in an adaptive setting.

5.1 Evaluation metrics

We argue that an anonymization mechanism deployed in real world applications should provide protection against advanced deanonymization attacks, preserve the core information of the original data (e.g., sentiment for product reviews), be semantically similar to the original sentences and of high quality in terms of language.

We measure the first two properties using both static (i.e., trained on source data) and adaptive (i.e., trained on data perturbed by the respective mechanism) BERT-based (Devlin et al., 2019) author and sentiment classifiers by fine-tuning the pretrained language model's top three layers and using a two-layer classifier for the author and sentiment labels. BERT has proven to be successful for both sentiment classification (Sun et al., 2019) and authorship attribution (Fabien et al., 2020) and thus represents a suitable model for both tasks. We report all classification results in terms of Matthews Correlation Coefficient (MCC) (Matthews, 1975; Gorodkin, 2004). An MCC score of +1 means perfect predictions whereas 0 indicates random guessing. MCC is more suitable to assess classification performance than accuracy (Chicco and Jurman, 2020) as it is not easily fooled by biased classifiers in case of imbalanced datasets.

To assess the trade-off between attack (authorship attribution) and utility (sentiment analysis), we measure each method's *relative gain* based on the original and obfuscated classification scores: Let A_o, S_o represent the MCC scores of the author and sentiment classifiers based on the original data, and similarly, let A_p, S_p represent the scores on perturbed data normalized to the range $[0, 1]$. Then we define its relative gain as $\gamma := S_p/S_o - A_p/A_o$.

To measure semantic similarity between the anonymized and original sentences, we compute the cosine similarity of their representations obtained by SBERT (Reimers and Gurevych, 2019), which is a model that has been optimized for capturing semantic similarity between textual inputs. For language quality, we compute the average perplexity (PPL) of the pretrained GPT-2 (Radford et al., 2019) over the output sentences of each model.

Table 1: Performance of authorship and sentiment classifiers trained and evaluated on data generated by anonymization mechanisms as measured by MCC scores. **Best trade-offs** are identified by the *relative gain* metric introduced in section Section 5.1

	original	GloVe embeddings				Poincaré embeddings				Paraphrase ($\epsilon = 1/T$)			
Privacy budget ϵ	∞	6	8	10	12	0.5	1	2	8	0.05	0.1	1.0	10
IMDb:													
Author (static)	0.98	0.12	0.20	0.28	**0.33**	0.87	0.88	0.88	0.89	0.19	0.21	0.22	**0.22**
Author (adapt.)	0.98	0.58	0.79	0.90	0.95	0.97	0.97	0.98	0.98	0.62	**0.63**	0.64	0.66
Sentim. (static)	0.71	0.21	0.32	0.43	**0.50**	0.53	0.52	0.52	0.53	0.37	0.40	0.40	**0.42**
Sentim. (adapt.)	0.71	0.22	0.37	0.52	0.60	0.56	0.54	0.56	0.56	0.40	**0.42**	0.41	0.43
SBERT CS	1.00	0.30	0.49	0.70	0.85	0.66	0.67	0.68	0.68	0.58	0.61	0.62	0.63
PPL	44.5	5003	3544	1414	512	431	384	330	310	37.2	34.8	34.4	33.9
Yelp:													
Author (static)	0.80	0.12	0.23	**0.40**	0.49	0.59	0.61	0.60	0.62	0.22	0.35	0.37	0.38
Author (adapt.)	0.80	0.32	0.47	0.62	0.68	0.72	0.73	0.73	0.75	**0.35**	0.35	0.37	0.39
Sentim. (static)	0.51	0.14	0.20	**0.27**	0.33	0.35	0.37	0.36	0.37	0.20	0.21	0.23	0.24
Sentim. (adapt.)	0.51	0.17	0.26	0.34	0.43	0.44	0.45	0.45	0.46	**0.32**	0.30	0.30	0.33
SBERT CS	1.00	0.29	0.43	0.60	0.76	0.35	0.37	0.38	0.38	0.49	0.51	0.54	0.54
PPL	99.7	13427	8555	3061	1534	1248	1232	1155	1116	148	143	138	132

5.2 Implementation Details

We implement both mechanisms proposed in the papers by Feyisetan et al. (2020, 2019) using numpy. Concretely, we use 50-dimensional GloVe (Pennington et al., 2014) vectors as our Euclidean embeddings and train 50-dimensional Poincaré embeddings on our own. For the latter, we extract ∼1,300,000 word tuples representing hypernymy relationships for IMDb and ∼1,800,000 tuples for Yelp from WebIsADB[3] (Seitner et al., 2016) by removing words with less than 10 occurrences and keeping only tuples contained in the GloVe vocabulary as well as the respective review dataset[4].

When encountering out-of-vocabulary (OOV) words, Algorithm 1 cannot assign embeddings and thus not perturb them, which violates DP. Also, merely removing the words does not change this as it changes the length of the output text while DP is only fulfilled for texts of the same length. For GloVe embeddings, a relevant effect in terms of experimental results is not present as the large vocabulary covers almost all words we encounter. The vocabulary size of our Poincaré embeddings is however limited (∼ 10,000) and, following Feyisetan et al. (2019), does not contain stopwords. As we aim to compare methods outputting human-readable texts and the removal of stopwords clearly affects readability, we instruct the mechanism to simply ignore OOV words. The results for removing OOV words can be found in Table 3 in the appendix.

For GPT-2 and BERT, we use the pretrained checkpoints from the HuggingFace transformers library (*gpt2, bert-base-uncased*; 117M, 110M parameters) and fine-tune each instance on a single NVIDIA T4 GPU.

5.3 Datasets

We conduct experiments using IMDb movie reviews (Maas et al., 2011) and Yelp business reviews[5] which contain author and sentiment labels in the form of ratings on the scale of 1-10 and 1-5, respectively. For both sources, we keep data from ten users with the most reviews, hereby obtaining dataset sizes of 10,000 for IMDb and 15,729 for Yelp. We simplify sentiment labels by rating movie reviews with ≥ 5 points and business reviews with ≥ 3 points as positive and the rest negative.

5.4 Results

Table 1 shows that paraphrasing significantly outperforms word-level mechanisms in terms of protection against adaptive adversaries. When evaluating privacy and utility for static classifiers, it becomes apparent that small perturbations are enough to trick author classifiers. Therefore, for static classifiers, mechanisms with weak word-level perturbations caused by smaller ϵ values show an equal trade-off on IMDb and a slightly better trade-off on Yelp reviews as they better preserve the sentiment

[3]Terms of use and license information: Appendix A.1

[4]As the procedure was not fully described in the paper, we increased (by factor ≥ 10) the training data of the original work, hereby having a larger vocabulary and more variation in the perturbations. We do so to minimize the risk of bad results merely due to implementation issues.

[5]https://www.yelp.com/dataset/

than the stronger changes caused by our model. Notably, paraphrasing shows better semantic preservation as well as higher language quality as measured by PPL when comparing it to word-level mechanisms calibrated for comparable privacy protection against the author classifier. This is also visible in the exemplary outputs provided in Table 2.

6 Related work

Other DP mechanisms for text Earlier mechanisms for differentially private text obfuscation settled for simpler output representations: Weggenmann and Kerschbaum (2018); Fernandes et al. (2019) employ Bag-of-Words (BoW) models and produce term-frequency vectors as output. Similarly, obfuscated dense vector representations are obtained in (Beigi et al., 2019) by perturbing the output of an encoder network. While not human-readable, these vector representations can be shared for automated processing, such as topic or sentiment inference and machine learning. To generate human-readable text, Bo et al. (2021) employ an encoder-decoder model similar to ours, but without paraphrasing, and sample output words using (a two-set variant of) the Exponential mechanism (McSherry and Talwar, 2007). Weggenmann et al. (2022) propose a differentially private variation of the variational autoencoder and use it as a sequence-to-sequence architecture for text anonymization.

Authorship obfuscation without DP Approaches not following DP range from rule-based algorithms relying on human-engineered text perturbations such as synonym replacements or word removals (Bevendorff et al., 2019; Mahmood et al., 2019) to methods incorporating deep learning. Models of the latter typically incorporate discriminator networks to penalize generating author-revealing information (Shetty et al., 2018; Xu et al., 2019). Similar to DP mechanisms, previous work is concerned with learning private text vector representations (Coavoux et al., 2018b).

Differentially private optimization Differentially private optimization algorithms such as DP-SGD and related methods (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016) have emerged as effective methods for protecting the training data of a model. Recent work has shown that both generative and discriminative language models can effectively be trained with these optimization approaches (Li et al., 2021; Yu et al., 2021) and therefore represent an important contribution for protecting against data leakage of language models (Song and Raghunathan, 2020; Carlini et al., 2021). These methods can be seen as complementary to the approaches discussed in this paper which protect data during inference.

7 Conclusion

We discussed and demonstrated the weaknesses of word level DP mechanisms and proposed a paraphrasing model circumventing most of these. We find that our approach outperforms word level mechanisms in terms of protection against adaptive adversaries, while the latter should be favored against weaker adversaries. Future work could address integrating auxiliary adversarial losses to paraphrasing systems or enabling paraphrases that better preserve the core information of the source text.

8 Ethical Considerations

Abuse of Anonymization Mechanisms Text Anonymization is an important field of research for the protection of privacy of individuals as well as for enabling freedom of speech. Still, anonymization mechanisms may be exploited for negative causes. Specifically, guaranteed anonymity on the internet might lead individuals to spread hate speech. Furthermore, mechanisms as ours can be used to anonymously generate and spread fake reviews or fake news. Important areas of research fighting these problems include hate speech and toxicity detection (Djuric et al., 2015; MacAvaney et al., 2019) as well as fake review detection (Mukherjee et al., 2013; Barbado et al., 2019) and fake news detection (Shu et al., 2017; Ruchansky et al., 2017).

Bias in Large Language Models Large language models such as GPT-2, which our proposed approach is based on, often inherit biases towards various demographics from the large amount of data they are trained on (Sheng et al., 2019; Abid et al., 2021). These biases can cause unforeseen effects when generating language output and could potentially alter statements of authors whose texts are being anonymized. An increasing amount of work is aiming to understand and tackle such biases in language models (Vig et al., 2020; Liang et al., 2021).

Evaluation Fairness In this paper, we evaluate our approach experimentally and compare its performance to mechanisms proposed in previous research works. Since no code was publicly released for the approaches we are comparing ours to, we implemented the mechanisms ourselves. While we replicated the original systems as close as possible to the description in the papers using all the information available, we cannot guarantee that they are exactly the same as not all the information about preprocessing and implementation details is publicly available.

Acknowledgements

We thank Zhijing Jin for the helpful discussions about the presentation of our results and the design of our paper.

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.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Mishari Almishari, Ekin Oguz, and Gene Tsudik. 2014. Fighting authorship linkability with crowdsourcing. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, page 69–82, New York, NY, USA. Association for Computing Machinery.

M. Andrés, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *ACM SIGSAC conference on Computer & communications security*, pages 901–914. ACM.

Rodrigo Barbado, Oscar Araque, and Carlos A Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. I am not what i write: Privacy preserving text representation learning.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108, Florence, Italy. Association for Computational Linguistics.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3997–4007, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3).

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg. Springer Berlin Heidelberg.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018a. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018b. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Damien Desfontaines. 2021. Local vs. central differential privacy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mael Fabien, Esaú Villatoro-Tello, Petr Motlícek, and Shantipriya Parida. 2020. Bertaa : Bert fine-tuning for authorship attribution. In *ICON*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

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. IEEE.

J. Gorodkin. 2004. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367–374.

Julian Hitschler, Esther van den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and POS-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58, Copenhagen, Denmark. Association for Computational Linguistics.

Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A. Hua. 2019. Syntactic recurrent neural network for authorship attribution. *CoRR*, abs/1902.09723.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264.

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 of *CEUR Workshop Proceedings*, pages 890–894. CEUR-WS.org.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94.

Fragkiskos Koufogiannis, Shuo Han, and George J Pappas. 2015. Optimality of the laplace mechanism in differential privacy. *arXiv preprint arXiv:1504.00065*.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

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.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71.

B. W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

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.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, et al. 2013. Fake review detection: Classification and analysis of real and pseudo reviews. *UIC-CS-03-2013. Technical Report*.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Zhan Qin, Yin Yang, Ting Yu, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2016. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 192–203.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Josyula R. Rao and Pankaj Rohatgi. 2000. Can pseudonymity really guarantee privacy? In *9th USENIX Security Symposium (USENIX Security 00)*, Denver, CO. USENIX Association.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. 2016. A large DataBase of hypernymy relations extracted from the web. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 360–367, Portorož, Slovenia. European Language Resources Association (ELRA).

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650.

Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Congzheng Song and Ananth Raghunathan. 2020. *Information Leakage in Embedding Models*, page 377–390. Association for Computing Machinery, New York, NY, USA.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 196–206, New York, NY, USA. Association for Computing Machinery.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Michael Tschuggnall and Günther Specht. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 195–199, Gothenburg, Sweden. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Benjamin Weggenmann and Florian Kerschbaum. 2018. *SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining*, page 305–314. Association for Computing Machinery, New York, NY, USA.

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.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Alexandra Wood, Micah Altman, Kobbi Nissim, and Salil Vadhan. 2020. Designing access with differential privacy. In Cole, Dhaliwal, Sautmann, and Vilhuber, editor, *Handbook on Using Administrative Data for Research and Evidence-based Policy*, chapter 6. wayne. Last accessed on 2022-01-13.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2020a. Differentially private adversarial robustness through randomized perturbations. *arXiv preprint arXiv:2009.12718*.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257.

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

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

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially private fine-tuning of language models.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Information about terms of use for data

In this section, we provide information and references about the terms of use and licenses of each dataset we are using.

WordNet Wordnet can be downloaded and accessed online without specifically requesting access and can be used for research and also commercial applications in accordance with the WordNet 3.0 license: https://wordnet.princeton.edu/license-and-commercial-use

WebIsADB WebIsADB can be downloaded and accessed online without specifically requesting access. The dataset is licensed under a Creative Commons Attribution-Non Commercial-Share Alike 3.0 License: http://creativecommons.org/licenses/by-nc-sa/3.0/

IMDb IMDb movie reviews can be downloaded and accessed online without specifically requesting access. Unfortunately, we could not find information about license specifications. More information is available at https://ai.stanford.edu/~amaas/data/sentiment/

Yelp Researchers aiming to use the Yelp dataset have to sign the terms of use (https://s3-media3.fl.yelpcdn.com/assets/srv0/engineering_pages/bea5c1e92bf3/assets/vendor/yelp-dataset-agreement.pdf). For commercial use, researchers should contact Yelp via dataset@yelp.com. More information is available at https://www.yelp.com/dataset.

Table 2: Exemplary output of anonymization mechanisms for Yelp data

Exemplary Reviews for Yelp
Original: This store is so adorable . In addition to baked goods they offer sandwiches for breakfast and lunch . The turkey sandwich was excellent . The textures were perfect though, especially for the almond amaretto cookie . It had the right balance of chewy with a slight amount of crunch.
Euclidean embedding perturbations ($\epsilon = 8$): designated store is work adorable making top tubular continue watered goods do offer salad ranging breakfast filling 5,000-a carries original turkey sandwich was excellent parts national textures were play never neighbors with for part mustard amaretto cookie hatred make had a direction balance end sugary another a erratic amounts of one-off today
Euclidean embedding perturbations ($\epsilon = 10$): fact store is 're granny his in health they dish goods kept offer sandwiches giving dinner besides lunch result the turkey sandwich given delivering . the textures ten captures . then especially own the apricot izola cookie on be had the right footing of chewy with a slight amount in crunch at
Poincaré embedding perturbations ($\epsilon = 1$): this flag is so adorable . in abundance to waffles chunk they many slimy for eggs and peppery . the vindaloo stickers was excellent . the blt were splurge gun , especially for the quail amaretto crunch . it had the quantity observation of crisp with a trotter simple of crunch .
Poincaré embedding perturbations ($\epsilon = 2$): this patient is so adorable . in stuff to asparagus walk-up they con jets for pricy and tamale . the petite cook was excellent . the rang were steal train , especially for the updated amaretto soak . it had the say many of containing with a slight land of cans .
Paraphrased ($\epsilon = 0.1$): There is a cute store. There is a sandwich being served by the sandwich shop. The sandwich is tasty. The two textures are alike. There were chews on the chem.
Paraphrased ($\epsilon = 1$): There's adorable store in this photo. In addition to baked goods they offer sandwiches for breakfast and lunch. This was a great sandwich. The desserts taste delicious! The food was chewy.

Table 3: Results for hyperbolic perturbations (Feyisetan et al., 2019) when removing out-of-vocabulary words.

	original	Poincaré embeddings			
Privacy budget ϵ	∞	0.5	1	2	8
IMDb:					
Author MCC (static)	0.98	0.03	0.12	0.07	0.12
Author MCC (adapt.)	0.98	0.67	0.69	0.68	0.69
Sentim. MCC (static)	0.71	0.27	0.30	0.31	0.28
Sentim. MCC (adapt.)	0.71	0.35	0.40	0.38	0.39
SBERT CS	1.00	0.32	0.33	0.33	0.34
Yelp:					
Author MCC (static)	0.80	0.14	0.14	0.15	0.16
Author MCC (adapt.)	0.80	0.32	0.35	0.34	0.36
Sentim. MCC (static)	0.51	0.17	0.18	0.20	0.20
Sentim. MCC (adapt.)	0.51	0.21	0.23	0.24	0.25
SBERT CS	1.00	0.54	0.54	0.56	0.57