Plausible Deniability of Redacted Text

Vaibhav Gusain[1 [0000−0002−7008−5201]] and Douglas Leith[1 [0000−0003−4056−4014]] ⋆

Trinity College Dublin, Ireland

Abstract. Providing privacy for natural language text data remains a largely open problem, despite its great practical importance. The current state of the art is manual redaction of sensitive words such as names, addresses etc. In this paper we propose viewing a corpus of text as a probability distribution over sequences of words. A sentence is then one realization from this distribution and redacting words changes the probability distribution. We use the Renyi-divergence divergence as a measure of the distance between two redacted datasets. We show that if enough words are redacted then sensitive redacted text can be made be statistically indistinguishable from non-sensitive redacted text. This can be used to develop efficient redaction strategies, that minimise the amount of redaction while meeting a privacy target.

Keywords: Data Privacy · Natural Language Processing · Text Sanitization.

1 Introduction

Training of neural nets such as large language models requires the availability of natural language text training data. In this paper we revisit the question of how to sanitise sensitive text data so that it can be used for model training while preserving privacy. We introduce a new approach for quantitatively estimating the privacy gain from text redaction and demonstrate its usefulness on a wide range of datasets. This can be used to develop efficient redaction strategies, that minimise the amount of redaction while meeting a privacy target. The approach is closely related to differential privacy, but differs in several respects that are important for text data.

There are two main approaches to enhancing privacy when training machine learning models. These approaches are complementary, and both can be used together. One is to add noise to the gradient updates used during training, e.g. see DP-SGD (1) (17) and related work. The other is to sanitise the training data itself. Here we focus on the latter.

When the training data is numeric then the addition of appropriate noise, e.g. Laplacian noise scaled proportionally to the differential privacy (DP) "sensitivity" of the data, can be used to enforce differential privacy guarantees (7).

⋆ This work was supported by Science Foundation Ireland grant 16/IA/4610.

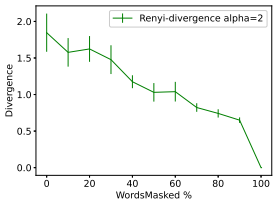


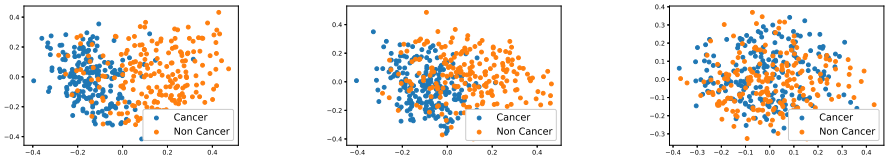
Fig. 1: Measured Renyi vs random redaction level for Medal dataset.

While it is tempting to apply this approach to text by mapping the text to a numeric embedding vector, adding noise, and then mapping back to text (8; 5; 20), this creates a host of unpleasant issues. For example, the nature of the mapping from text to vectors (which texts are mapped to vectors near or far away from one another) directly affects the impact of added noise, and so privacy. However, this is poorly understood, especially for modern embedding approaches based on neural networks. Another deeply problematic issue is that the words in a sentence tend to be correlated in complex ways, making any privacy approach based on individual words tend to overestimate the privacy gained.

In practice, the most popular approach for sanitising text data is redaction i.e. replacing selected words with an uninformative mask token. This is, for example, already widely used to remove personally identifying information (PII), e.g. names and addresses, from documents (12). However, other aspects of the text data can also be sensitive. For example, the text data may reveal sexual/gender traits, political preferences, health-related information/concerns, social and racial characteristics, etc of the user population from which the data was gathered. Textual style can also act as a personal fingerprint facilitating de-anonymisation and membership attacks against neural nets trained on this data.

Protecting privacy is especially challenging with text data because simply redacting specified keywords is rarely enough: the surrounding context can easily continue to reveal sensitive information (3). For example, in the sentence "I am diagnosed with cancer. I have to go to St Lukes for chemotherapy and will probably lose my hair" redacting "cancer" and "chemo" is not sufficient to conceal the cancer diagnosis if St Lukes is known to be a cancer care hospital. If "St Lukes" is also redacted, the combination of "diagnosis" and "lose my hair" is still enough to indicate a cancer diagnosis with high probability.

In summary, there is an urgent need more effective methods for improving the privacy of text data. Given the challenging nature of natural language text, it is probably too much to hope for theoretical guarantees but that should not stop us from trying to develop useful methods motivated by theoretical analysis. In this paper we take a step in that direction. We use redaction to add "noise" to text and by staying within original text domain thereby avoid most of the issues with numerical embeddings, and by working with text corpuses rather



(a) No redaction (b) 20% redaction (c) 40% redaction

Fig. 2: Illustrating the increasing overlap between the semtence embeddings for cancer and non-cancer text from the Medal dataset as the level of redaction is increased. SentenceBERT embeddings are projected to two dimensions using PCA, random redaction is used.

than individual words or sentences we can accommodate the word correlations within sentences.

1.1 Our approach

A dataset $\mathcal{D}$ is a collection of items. Each item is a sequence $x = (x_1, ..., x_{|x|})$ of words x_t belonging to a fixed vocabulary and with length $|x| \leq N$, N being the maximum admissible length. A redaction policy $\pi_p(x)$ maps sequence x to a new sequence where some words have been redacted i.e. replaced by an uninformative mask token `MASK`. We will assume that every redaction policy is parameterised by a parameter p taking a value between 0 and 1 such that when $p = 0$ then no words are redacted, when $p = 1$ then every word is redacted. For example, the uniform random $\pi_{rand,p}(x)$ redaction policy redacts each word in sequence x with probability p. Alternatively, we might rank the words in our vocabulary by their sensitivity and redact the top p fraction of these.

Each item x in a dataset is a random draw from a probability distribution $P(x)$ over sequences of words. After redaction, each element x is mapped to a new sequence $redact(x)$ and the redacted dataset becomes a sample from probability distribution $redact(P)$. We measure the distance between two redacted datasets $redact(\mathcal{D}_0)$ and $redact(\mathcal{D}_1)$ by the smallest value of $\epsilon \geq 0$ such that $\tilde{P}_0(y) \leq e^{\epsilon}\tilde{P}_1(y) + \delta$ and $\tilde{P}_1(y) \leq e^{\epsilon}\tilde{P}_0(y) + \delta$ where $\tilde{P}_0 := redact(P_0)$ is the probability distribution over token sequences in dataset $redact(\mathcal{D}_0)$, $\tilde{P}_1 = redact(P_1)$ in dataset $redact(\mathcal{D}_1)$ and y is any redacted sequence of words with length $|y| \leq N$.

This distance measure is similar to that used in (ϵ, δ)-differential privacy but with the difference that the set of neighbouring databases now consists of the single database $redact(\mathcal{D}_0)$ rather than all databases differing from $redact(\mathcal{D}_1)$ by a single element. When ϵ, δ are sufficiently small, the publication of private dataset $redact(\mathcal{D}_1)$ then only provides an attacker with limited new information over and above that already available from the public dataset $\mathcal{D}_0$. That is, we gain privacy in the sense of *indistinguishability* between the $redact(\mathcal{D}_0)$ and $redact(\mathcal{D}_1)$ datasets. It will prove convenient to work in terms of the Renyi-divergence

$D_\alpha(\tilde{P}_0||\tilde{P}_1)$ to calculate the distance between the datasets. We then convert this to an (ϵ, δ)-privacy guarantee using equations (2) and (3). See Section-4.1 and Section-4.2 for more details.

We assume the availability of a "safe" dataset $\mathcal{D}_0$ e.g. a public dataset that is suitably diverse and non-sensitive. Applying redaction policy π_p to both the sensitive dataset $\mathcal{D}_1$ and the safe dataset $\mathcal{D}_0$ then we expect that distance ϵ between the datasets decreases as the level of redaction increases, the distance becoming zero after redaction with $p = 1$.

Figure 1 illustrates this for the Medal dataset of medical records (see below for further details). The original dataset is split into a dataset $\mathcal{D}_1$ of cancer patients and a dataset $\mathcal{D}_0$ of non-cancer patients. Random redaction is used. The figure shows the measured Renyi-Divergence $D_{max}(P_0||P_1)$ between the empirical probability distributions P_0 and P_1 induced by $\mathcal{D}_0$ and $\mathcal{D}_1$ as the level of redaction is varied. As expected, it can be seen that the divergence decreases as the amount of redaction increases i.e. the two datasets become more similar.

Figure 2 illustrates this behaviour more visually. Redacted sentences are mapped to embedding vectors using SentenceBERT (15), the vectors are then projected onto to dimensions using PCA and shown as a scatter plot. It can be seen that without redaction the sentence embeddings have little overlap but as the level of redaction increases the overlap between the embeddings increases, indicating that distinguishing between the two datasets is becoming harder.

We make the following observations. (i) Redaction sanitizes the text data itself (rather then vector embeddings) and so yields a sanitized dataset that can be used for training ML models that take word sequences as input. (ii) By working in terms of the probability distribution over word *sequences* we take account of the correlation between the words in a sentence (the word context) and the associated potential for leakage of sensitive information. (iii) Sanitising the dataset is akin to local differential privacy i.e. the input to a query is perturbed to ensure privacy rather than the output of the query being perturbed. (iv) As the distance ϵ between the sanitized dataset and the safe dataset decreases, privacy increases but of course we expect utility to decrease (the added value of the new dataset decreases).

2 Related work

The existing literature on enhancing the privacy of text data can be roughly categorised as follows:

Redacting PII. Much of the literature on text redaction has focussed on redacting personally identifying information (PII). For example, (12) uses an ensemble of deep learning methods to detect and redact PII information from the medical notes of the patient, (2) considers the discovery of names, home towns, etc in student discussion boards. Other recent work includes (6) (21) (16).

Word-Level DP. Word-level DP approaches map an individual word to a vector embedding, add noise and then either map back to a new word or use the

noisy embedding directly. See e.g. (8; 20; 5). A typical choice of vector embedding is Glove (14). The choice of vector embedding has to made up front and its properties affect the privacy gained[1] in ways that remain very poorly understood. Words are discrete quantities and the impact of quantisation when mapping from vectors back to words also remains poorly understood. The DP "database" is a sentence and the words are the database entries. The DP guarantee (modulo concerns regarding the embedding already noted) therefore relates to insensitivity to an individual word in a sentence. However this DP analysis ignores correlations between the words in a sentence and so can underestimate the information release. The impact of correlations on DP is well known and was first noted by (9). See (10) for further discussion on the deficiencies of word-level DP.

3 Threat model

We consider the use of natural language text datasets for training machine learning models. We assume that the sensitive dataset itself is stored securely, but the sanitized/redacted dataset is publicly released. The main threat we consider is that that the sanitized dataset may be used to infer sensitive user traits e.g sexuality, health conditions, political preferences. This is a particularly topical concern since the development of LLMs is currently being driven by companies with a commercial interest in identifying user traits e.g for use in targeting adverts or other services. We assume that PII (names, addresses etc) has been removed, there being many existing techniques for this, see e.g. (6) (21) (16)

4 Preliminaries

The Renyi-divergence of order $\alpha > 1$ between two probability distributions P_0 and P_1 on sample space Y is (13):

$$D_\alpha(P_0||P_1) = \frac{1}{\alpha - 1} \log \int_Y P_0(x)^\alpha P_1(x)^{1-\alpha} dx \tag{1}$$

and similarly for $D_\alpha(P_1||P_0)$. When $\alpha = 1$ the Renyi-divergence equals the KL-divergence (11).

We say that two probability distributions P_0 and P_1 are (ξ, ρ)-zero-concentrated differentially private when:

$$D_\alpha(P_0\|P_1) \leq \xi + \rho\alpha \tag{2}$$

for all $\alpha > 1$. In the differential privacy literature P_0, respectively P_1, is the probability distribution induced by a randomised mechanism $\mathcal{M}$ applied to dataset

[1] Adding noise to an embedding perturbs it to nearby words, the way in which words are mapped to be close together (or far apart) therefore directly affects the output of the word-level DP sanitisation process.

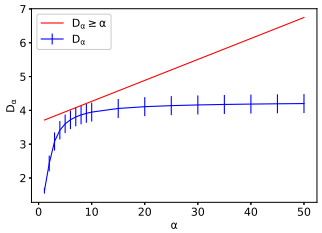


Fig. 3: Divergence vs α for non-redacted cancer and non-cancer text from Medal medical dataset.

$\mathcal{D}_0$, respectively $\mathcal{D}_1$ (4). Inequality (2) is then required to hold for all neighbouring datasets $\mathcal{D}_0$, $\mathcal{D}_1$, where datasets are neighbours if they differ by a single element. However, here we will consider other choices for P_0 and P_1. Specifically, they will be the distributions from which two datasets $\mathcal{D}_0$ and $\mathcal{D}_1$ are sampled.

When P_0 and P_1 are (ξ, ρ)-zCDP then from the proof of Lemma 21 in (4),

$$P_1(x) \leq \exp(\epsilon) P_0(x) + \delta \tag{3}$$

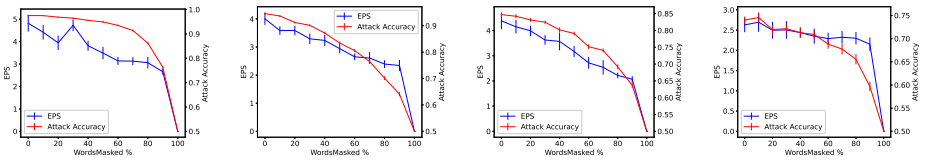
for every $\delta > 0$ where $\epsilon = \xi + \rho + 2\sqrt{\rho \log \frac{1}{\delta}}$. We will use (3) to map from Renyi-divergence curves to an (ϵ, δ) privacy guarantee.

4.1 Estimating Renyi-Divergence

To estimate the Renyi-divergence between two datasets we extend the estimator of (13), which is observed to scale well for high dimensional data. We updated the estimator to handle duplicate word-sequences, since these can become common following redaction[2]. In addition, each word sequence is mapped to a vector embedding[3] X_i. These are then fed to the estimator to calculate the Renyi-divergence. We use boot-strapping to calculate confidence intervals for the estimate. Namely, we sample with replacement n times from $redact(\mathcal{D}_0)$ and $redact(\mathcal{D}_1)$, estimate $D_\alpha(P_0||P_1)$ is calculated for each sample and then the mean and standard deviation of these n estimates calculated. We select n by calculating the mean and standard deviation vs n and selecting a value large enough that these are convergent. The mean of the estimated Renyi divergence is shown in our plots with the standard deviation indicated by error bars.

[2] See Appendix https://anonymous.4open.science/r/appendix_repo-F4CC for more details.

[3] The choice of embedding will, in general, affect the estimated divergence. This can be mitigated by calculating the divergence for many different embeddings and using the worst-case (i.e. largest) value. However, we found the impact to be relatively minor in practice, see Section-6.3, and SentenceBERT (15) to work well.



(a) Medal dataset (b) Political dataset (c) Amazon dataset (d) Reddit dataset

Fig. 4: Measured ϵ between redacted sensitive and safe datasets vs redaction level; random redaction. A lower value indicates better privacy. Also shown is the measured accuracy of a classification attack that tries to label which dataset the redacted sensitive text originated from (lower accuracy therefore equals greater privacy, with a classification accuracy of 50% corresponding to a random classifier).

4.2 Calculating (ϵ,δ)

To calculate ξ and ρ in equation (2), we first calculate D_α for a range of α values[4]. We then find a line that lies above the D_α vs α curve and select ρ as the slope of the line and ξ the intercept. Of course, many lines lie above the D_α curve, so we try to select one such that ρ and ξ are minimised. See for example Figure-3, which shows D_α vs α for the Medal medical dataset (the blue curve). This curve is upper bounded by the red line. The values of ρ and ξ corresponding to this red line are plugged into equation (3 to obtain the corresponding (ϵ,δ) privacy values.

Note. We use $\delta = 0.00008$ in all of our experiments as it is encouraged to keep $\delta < \frac{1}{n^2}$ where n is the number of input points (1).

5 Experiments

5.1 Datasets

We evaluate performance using the following datasets, each of which we split into "sensitive" and "safe" datasets.

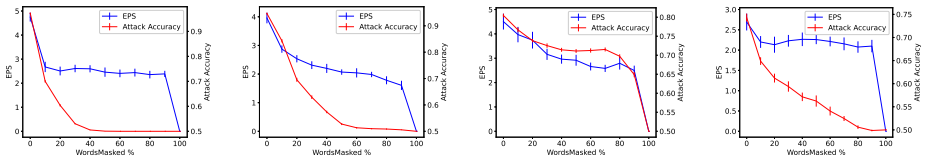
(i) Medal dataset (19)[5]. This dataset contains abstracts of medical papers, along with the diseases the abstract talks about. We partition this dataset into text with cancerous and non-cancerous diseases. Each dataset contains 2200 sentences. For our experiments, text with cancerous diseases was chosen to be the sensitive dataset.

(ii) Political dataset-(18)[6]. This contains comments on Facebook posts from 412 members of the United States Senate and House. Each comment is labeled

[4] We select the range to be large enough that D_α no longer increases as we increase α.

[5] https://huggingface.co/datasets/medal

[6] Data can be downloaded by following the instructions in the repository https://github.com/xuqiongkai/PATR



(a) Medal dataset (b) Political dataset (c) Amazon dataset (d) Reddit dataset

Fig. 5: Measured ϵ between redacted sensitive and safe datasets vs redaction level; more efficient redaction strategy. A lower value indicates better privacy. Also shown is the measured accuracy of a classification attack that tries to label which dataset the redacted sensitive text originated from (lower accuracy therefore equals greater privacy, with a classification accuracy of 50% corresponding to a random classifier).

with the corresponding Congressperson's party affiliation i.e. S ϵ {democratic, republican } We partition the dataset into text from users with Republican and Democrat political preferences. Each dataset contains 2000 sentences. For our experiments, text from users with Republican political preferences is chosen to be the sensitive dataset.

(iii) Amazon dataset[7]. This dataset contains product reviews from Amazon customers. We selected the reviews which were categorised as "drug-store" and "kitchen-appliances". For our experiments, the dataset with drug-store reviews was chosen to be the sensitive dataset.

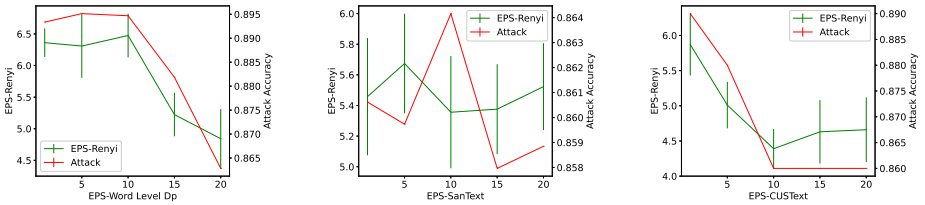
(iv) Reddit dataset[8]. This dataset contains post content from the subreddits r/depression and r/SuicideWatch. We partition this data into posts related to suicide and depression. Each dataset contains 2000 sentences. For our experiments, the text from the suicide subreddit was chosen to be the sensitive dataset.

5.2 Enhancing Privacy By Redaction

For each of the datasets we measured the Renyi-divergence as the percentage of words redacted using a random redaction policy was varied from 0 to 100%. The divergence values were then converted to ϵ values as explained previously. The results are shown in Figure 4.

To help gain confidence that redaction really is improving privacy, we carry out a simple classification attack. The redacted datasets are split into training and test data (90:10 split). A classifier is trained on this data, taking a sequence of words as input and outputting an estimate of whether the sentence came from the safe or sensitive datasets. Since there are only two classes and the data is balanced, a classification accuracy of 50% corresponds to a random classifier.

[7] https://huggingface.co/datasets/amazon_reviews_multi
[8] https://www.kaggle.com/general/256134



(a) Word-level DP (b) San-Text (c) CusText

Fig. 6: Measured ϵ-renyi and attack accuracy for various word-level DP approaches applied to Medal Dataset.

Figure 4 shows how the measured accuracy of this classifier varies with the redaction level for each of the datasets. It can be seen that the accuracy decreases as the redaction level increases, and that this decrease is roughly proportional to the decrease in the measured ϵ value.

5.3 Smarter Redaction

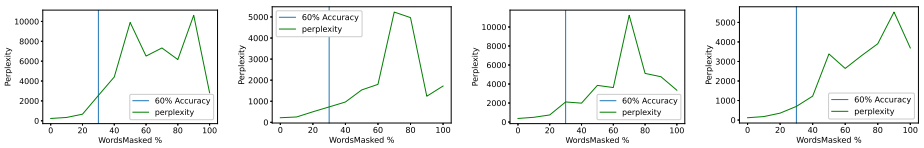
It can be seen from Figure 4 that when random redaction is used then relatively high levels of redaction are needed to ensure smaller ϵ values. Of course random redaction is rather crude, and smarter redaction approaches (in the sense that they achieve a target ϵ with fewer words redacted) are certainly possible.

We illustrate the scope for smarter redaction via a simple approach based on logistic regression weights. Namely, we took the logistic regression classifier from Section 5.2 and ranked the words from the datasets by the magnitude of the weight assigned to them by this classifier. We then redact the top p percent of these words from the datasets when redacting at level p.

Figure 5 plots the measured ϵ between the sensitive and safe datasets as the redaction level is increased in this way. It can be seen that a much lower level of redaction is now needed, compared to Figure 4, to obtain a given ϵ value. Also shown in Figure 5 is the measured accuracy of the simple classification attack as the redaction level is varied and it can be seen that the accuracy also now falls much more rapidly. For example, in Figure 4(a) a redaction level of around 90% is needed to reduce the accuracy of the attack to 60% (recall an accuracy of 50% corresponds to a random classifier, so 60% represents a high level of privacy), for the Medal dataset while with the smarter redaction strategy a redaction level of around 30% is sufficient to achieve this. Observe also that there is now a clear "knee" in the measured divergence vs redaction level curve, with the knee corresponding a low attack accuracy. It can be seen from Figure 5 that the behaviour is also similar for the other datasets studied.

5.4 Word-level DP

Word-level DP approaches sanitize text by converting each individual word to a vector embedding, adding noise to the embedding, and then mapping the noisy



(a) Medal dataset (b) Political Dataset (c) Amazon Dataset (d) Reddit dataset

Fig. 7: Measuring impact of privacy on utility. Next word prediction performance for LSTM trained on redacted dataset. Performance is measured on non-redacted data. The vertical line indicates the redaction level that reduces classification attack accuracy to 60%, taken from Figure 5.

embedding back to a word (8; 20; 5). In this section, we compare our approach to word-level DP approaches.

As discussed previously, word-level DP approaches aim to hide the information revealed by the individual words and so can fail to hide information revealed by the sentence as a whole[9].

We illustrate this by conducting the same attack as before on the word-level DP sanitized data, while also checking the ϵ (indicated by ϵ-renyi) between the sensitive and safe datasets. A high attack accuracy indicates that sensitive information is leaked from the sanitized sentences. Similarly, a high ϵ-renyi indicates that there are significant differences between sensitive and non-sensitive datasets.

Figure 8 shows the measured ϵ-renyi for the Medal dataset as the level of noise is increased (indicated by ϵ) for various word-level DP approaches. Also shown is the measured accuracy of our classification attack. It can be seen that even when a great deal of noise is added (low ϵ values), both the ϵ-renyi values and the attack accuracy remain high. Results for other datasets show similar behaviour and are provided in the Appendix [10].

5.5 Utility vs Privacy

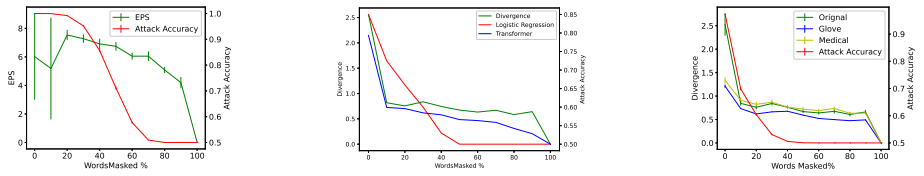
In general, we expect there to be a trade-off between privacy and utility. By redacting text to make a sensitive training dataset more private, the value of the sensitive training data is likely to be reduced because (i) redaction reduces the textual information contained in the sensitive dataset and (ii) by making the sensitive dataset more similar to an existing safe public dataset the added value over only using the public data for training is reduced.

To investigate this trade-off we trained a next-word-prediction model using PyTorch. A standard LSTM-RNN model[11] was used with two layers, each layer

[9] In particular, the DP analysis ignores correlations between the words in a sentence and so may greatly underestimate the information release. The impact of correlations on DP is well known and was first noted by (9).

[10] https://anonymous.4open.science/r/appendix_repo-F4CC

[11] Training code can be found at: https://github.com/pytorch/examples/tree/main/word_language_model



(a) (b) (c)

Fig. 8: 8a Measured ϵ and attack accuracy for cancer sentences when compared against IMDB reviews.8b Measured Renyi-divergence ($\alpha = 2$) and attack accuracy for logistic regression and BERT transformer classification attacks as the redaction level is increased. Medal dataset. 8c Measured Renyi-divergence ($\alpha = 2$) with different embeddings: (i) general-purpose sentenceBERT, (ii) fine-tuned medical sentenceBERT, (ii) Glove. Medal dataset.

having 200 hidden states and an input Embedding layer. The model has 4,041,675 parameters. A dictionary of all words was created from the dataset and input text was vectorized by replacing each word with its corresponding index in the dictionary. The model was trained using a negative log-likelihood loss.

The sensitive dataset was divided into a train-validation (90:10) split. The training data was redacted while validation data was not redacted. The model was then trained on the redacted training data and was tested against the held-out validation data. The model performance was evaluated by the measured perplexity (ppl) on the validation set, the lower the value of perplexity the better the model is at predicting a given sequence.

Figure 7 shows the measured perplexity on the validation data of the next word prediction model for each of the datasets studied as the redaction level is varied (for the smarter redaction approach of Section 5.3). The vertical line in the plot indicates the point where the measured attack classification accuracy is 60% (taken from Figure 5). It can be seen as expected, the perplexity increases as the redaction level increases. It can be seen that the perplexity increases with the level of redaction, as expected. However, for redaction levels up to 20-30% the perplexity of the model trained on the redacted dataset remains fairly close to the perplexity of the model trained on non-redacted dataset. A redaction level of 30% is sufficient to bring the attack classification accuracy down to 60% i.e a good degree of privacy can be obtained while preserving the utility of the dataset for model training.

6 Discussion

6.1 Choice of the Safe dataset

When the safe dataset is similar to the sensitive dataset, then we can expect that only a small amount of redaction is needed to bring the two datasets closer

together. Conversely, we expect that a higher level of redaction is needed to make very disparate datasets similar.

This is illustrated in Figure-8a, which can be directly compared with Figure-5a. In both cases the Medal cancer text is the sensitive dataset but in Figure-8a the safe dataset is IMDB review text while in Figure-5a it is Medal non-cancer text. It can be seen that with the IMDB data almost 80% of the words need to be redacted to get an attack accuracy close to 50% whereas with the Medal non-cancer text a redaction level of around 50% is sufficient.

The choice of safe dataset is therefore a privacy design parameter that can be used to manage the trade-off between privacy and utility in a fairly transparent manner.

6.2 More Powerful Attacks

It is important to stress that it is the Renyi-divergence that provides a sound measure of privacy, not the accuracy of any specific attack. In the previous sections we use a classification attack based on logistic regression to roughly verify privacy. However, this attack on its own does not provide any privacy guarantee.

For example Figure-8b shows the attack accuracy and Renyi-divergence as the level of redaction is varied. It can be seen that at redaction levels around 50-80% the attack accuracy is low (close to 50%, the accuracy of a random classifier). However, the Renyi-divergence remains relatively high at around 2.5 at these redaction levels, indicating that the datasets remain dissimilar. We therefore trained a more powerful BERT transformer model and used it to carry out the classification attack. Figure-8b, shows the measured attack accuracy. It can be seen that at 50-80% redaction the attack accuracy now remains relatively high, demonstrating the predictive power of the Renyi-divergence approach.

6.3 Choice of Embedding

The estimator in Section 4 maps text to a vector embedding and then estimates the Renyi-divergence between sets of vectors. It therefore depends on the choice of embedding used. It is problematic for a privacy approach to depend on the choice of embedding since (i) the properties of these embeddings remain poorly understood and (ii) an attacker can easily use a different embedding. For example, if a general purpose embedding is used in the Renyi-divergence but the attacker uses a domain specific embedding, a natural concern is that attacker may be able to extract information even when the Renyi-divergence estimate is low.

One of the great advantages of redaction is that it does not depend on the choice of embedding, but rather works directly with the text data[12]. We can then

[12] And one of the major deficiencies of all approaches tied to a single up front choice of embedding, such as word-level DP approaches.

calculate Renyi-divergence estimates for different choices of embeddings and use the largest value to evaluate privacy.

For example Figure-8c shows the measured Renyi-divergence estimates for three different choices of embedding: (i) a general-purpose pre-trained sentence-BERT embedding [13], (ii) sentenceBERT after fine-tuning on medical data and (iii) Glove[14] (14). sentenceBERT is a state of the art transformer embedding, Glove is an older embedding commonly used on word-level DP.

It can be seen from Figure-8c that the Renyi-divergence estimates for the two sentenceBERT embeddings are almost the same and consistently higher than the Renyi-divergence estimate using Glove i.e. Glove overestimates privacy. The consistency in the divergence estimates between the general purpose and fine-tuned sentenceBERT embedding indicates the robustness of the general purpose sentenceBERT and is why we use it in our earlier plots.

6.4 Limitations

Renyi-divergence is an estimate. Probably the main limitation of our approach is that it uses an estimate of the Renyi-divergence rather than the true value. We partially mitigate this by also estimating confidence intervals. However use of an estimate seems unavoidable since the true divergence cannot be calculated for realistic text data, and for similar reasons theoretical differential privacy guarantees are intractable. We argue that adopting a pragmatic approach and using estimates provides a way forward that is both useful and represents significant progress over the state of the art in text privacy. This is particularly pressing given the prevalence of text data and the current great interest in using it to train large language models.

7 Conclusions

We revisit the question of how to sanitise sensitive text data so that it can be used for model training while preserving privacy. The great majority of the existing literature on privacy enhanced model training gains privacy by adding noise to gradients used for training. The amount of noise added needs to scale with the number of model parameters since the DP sensitivity scales with this. When the number of model parameters is large (as it usually is for language models), the amount of noise needed is considerable and adversely affects model utility. Sampling of the input data can be used to boost privacy, but requires effective anonymisation which can be hard to achieve in practice and reduces the volume of training data available. The data sanitisation approach that we consider complements this line of work and offers new ways to manage the trade-off between privacy and utlity.

[13] https://www.sbert.net/

[14] The embedding vector of each word in a sentence is calculated, and the mean of these vectors is used as the sentence embedding.

Bibliography

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. p. 308–318. CCS '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2976749.2978318, https://doi.org/10.1145/2976749.2978318

[2] Bosch, N., Crues, R., Shaik, N., Paquette, L.: " hello,[redacted]": Protecting student privacy in analyses of online discussion forums. Grantee Submission (2020)

[3] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., Tramèr, F.: What does it mean for a language model to preserve privacy? In: 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 2280–2292. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3534642, https://doi.org/10.1145/3531146.3534642

[4] Bun, M., Steinke, T.: Concentrated differential privacy: Simplifications, extensions, and lower bounds (2016)

[5] Chen, S., Mo, F., Wang, Y., Chen, C., Nie, J.Y., Wang, C., Cui, J.: A customized text sanitization mechanism with differential privacy. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 5747–5758. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.355, https://aclanthology.org/2023.findings-acl.355

[6] Doudalis, S., Kotsogiannis, I., Haney, S., Machanavajjhala, A., Mehrotra, S.: One-sided differential privacy (2017)

[7] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

[8] Feyisetan, O., Balle, B., Drake, T., Diethe, T.: Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In: Proceedings of the 13th International Conference on Web Search and Data Mining. p. 178–186. WSDM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3336191.3371856, https://doi.org/10.1145/3336191.3371856

[9] Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. p. 193–204. SIGMOD '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1989323.1989345, https://doi.org/10.1145/1989323.1989345

[10] Mattern, J., Weggenmann, B., Kerschbaum, F.: The limits of word level differential privacy. In: Findings of the Association

for Computational Linguistics: NAACL 2022. pp. 867–881. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.65, https://aclanthology.org/2022.findings-naacl.65

[11] Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF). IEEE (Aug 2017). https://doi.org/10.1109/csf.2017.11, http://dx.doi.org/10.1109/CSF.2017.11

[12] Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J.R., Ross, J.L., William A. Faubion, J., Halamka, J.D., Soundararajan, V., Ardhanari, S.: Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. medRxiv (2021). https://doi.org/10.1101/2020.12.22.20248270, https://www.medrxiv.org/content/early/2021/02/23/2020.12.22.20248270

[13] Noshad, M., Moon, K.R., Sekeh, S.Y., Hero, A.O.: Direct estimation of information divergence using nearest neighbor ratios. In: 2017 IEEE International Symposium on Information Theory (ISIT). IEEE (Jun 2017). https://doi.org/10.1109/isit.2017.8006659, http://dx.doi.org/10.1109/ISIT.2017.8006659

[14] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162

[15] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)

[16] Shi, W., Shea, R., Chen, S., Zhang, C., Jia, R., Yu, Z.: Just fine-tune twice: Selective differential privacy for large language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 6327–6340. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.emnlp-main.425

[17] Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. p. 1310–1321. CCS '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2810103.2813687, https://doi.org/10.1145/2810103.2813687

[18] Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., Tsvetkov, Y.: RtGender: A corpus for studying differential responses to gender. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1445

[19] Wen, Z., Lu, X.H., Reddy, S.: MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Association for Computational Lin-

guistics (2020). https://doi.org/10.18653/v1/2020.clinicalnlp-1.15, https://doi.org/10.18653%2Fv1%2F2020.clinicalnlp-1.15

[20] Yue, X., Du, M., Wang, T., Li, Y., Sun, H., Chow, S.S.M.: Differential privacy for text analytics via natural text sanitization. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3853–3866. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.337, https://aclanthology.org/2021.findings-acl.337

[21] Zhao, X., Li, L., Wang, Y.X.: Provably confidential language modelling. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 943–955. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.69, https://aclanthology.org/2022.naacl-main.69