

# Text Redaction for Privacy Preserving Model Training- Appendix.

Vaibhav Gusain<sup>1</sup> and Douglas Leith<sup>1</sup>

School of Computer Science and Statistics, Trinity College Dublin  
(gusainv,doug.leith)tcd.ie

## 1 Hyperparameters for Next Word Prediction

To train the Next-word-prediction model negative log-likelihood loss was used. The model's weights were initialized with a value drawn from a uniform distribution over  $(-0.1, 0.1)$ . A stochastic gradient descent algorithm was used. Each model was trained for at most 20 epochs, with an initial learning rate of 20. the learning rate was annealed by a factor of 4 if the validation loss between two epochs does not decrease.

## 2 Hyperparameters for DP-SGD

When training models with DP-SGD, we used  $\delta = 0.00008$  and the max-gradient clip norm  $C = 1.5$ .

## 3 Overview of the redaction pipeline

Figure-7, provides an overview of the redaction pipeline. A classifier is trained on on a held-out training data, taking a sequence of words as input and outputting an estimate of whether the sentence came from the safe or sensitive datasets. The trained model is then used to rank the words that needs to be redacted from the sentences present in the safe and sensitive datasets. Words are redacted according to their ranks. The redacted sentences are then sent to a sentenceBERT to generate the embeddings. The generated embeddings are then sent to a Renyi-divergence estimator which estimates the Renyi-divergence between the safe and sensitive dataset. The estimated divergence is then converted to a Differential privacy guarantee.

## 4 Renyi-Divergence

### 4.1 Renyi-Divergence estimator

The Renyi-divergence of order  $\alpha$  between two probability distributions  $P_0$  and  $P_1$  on sample space  $Y$  is given by ) equation 1 (in the paper). We need to accomodate probability distributions with both discrete and continuous parts since

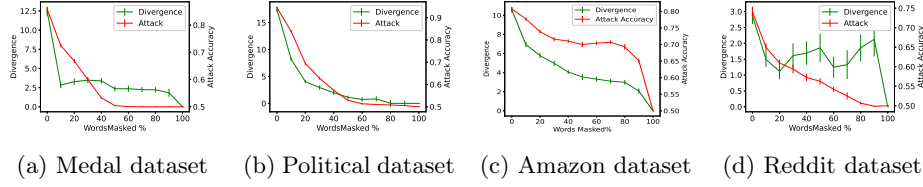


Fig. 1: Measured KL divergence between redacted sensitive and safe datasets vs redaction level; more efficient redaction strategy. Also shown is the measured accuracy of a classification attack that tries to label which dataset the redacted sensitive text originated from.

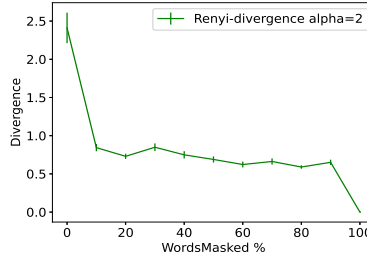


Fig. 2: Measured Renyi divergence for ( $\alpha = 2$ ) vs redaction level for Medal dataset.

after redaction there are often duplicate word sequences within a corpus of text (in the extreme case where all words are redacted every sentence becomes the same MASK token), corresponding to a discrete part in the probability distribution. We therefore decompose  $P_0$  into continuous pdf  $p_0$  and discrete probability mass function  $Q_0$ , similarly  $P_1$ , and then use a Monte Carlo plug-in estimator based on equation defined above.

In more detail, we draw  $N_0$  and  $N_1$  word sequences from datasets  $D_0$  and  $D_1$  respectively. We assume that these datasets are a representative sample from the probability distributions  $P_0$  and  $P_1$  respectively. We then map the  $i$ 'th word sequence to a vector embedding<sup>1</sup>  $X_i$  and round the elements of this vector to cluster duplicates. Letting  $Y_0 = \{X_i\}$  be the resulting multiset of vectors from  $D_0$  (a multiset can include duplicates) and  $Y_0^u \subset Y_0$  be the set of unique vectors in  $Y_0$ , we then estimate  $D(P_0||P_1)$  using

$$\hat{D}_\alpha(P_0||P_1) = \frac{1}{\alpha - 1} \sum_{y \in Y_0} \frac{1}{N_0} \log \frac{(\hat{P}_0(y))^\alpha}{\hat{P}_1(y)^{(\alpha - 1)}}$$

<sup>1</sup> 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 and SentenceBERT [5]

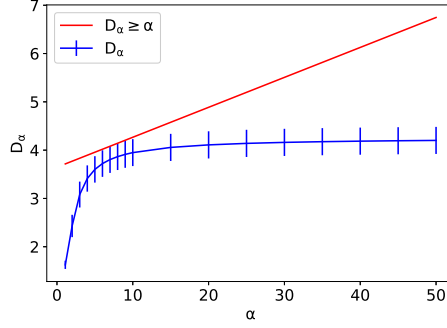


Fig. 3: Divergence vs  $\alpha$  for non-redacted cancer and non-cancer text from Medial medical dataset.

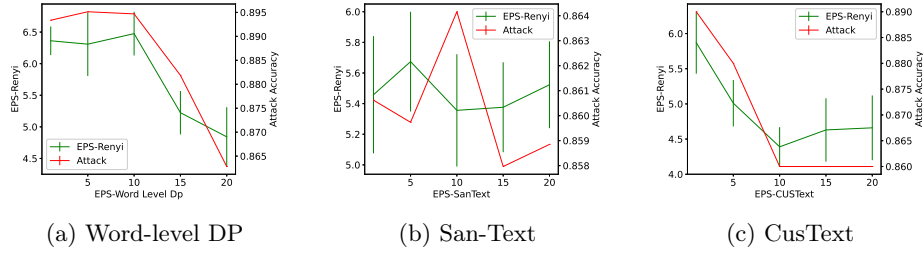


Fig. 4: Measured  $\epsilon$ -renyi and attack accuracy for various word-level DP approaches applied to Medial Dataset.

with

$$\hat{P}_0(y) = \frac{n(y, Y_0^u)}{N'_0 \rho(y, Y_0^u)^d}, \hat{P}_1(y) = \frac{n(y, Y_1^u)}{N'_1 \rho(y, Y_1^u)^d}$$

where  $n(y, Y^u) = \sum_{x \in N_k(y, Y^u)} |x|$  counts the nearest neighbours of  $y$  in set  $Y^u$ ,  $N_k(y, Y^u)$  is the set of  $k$ 'th nearest neighbors to  $y$  in set  $Y^u$  and  $|x|$  is the multiplicity of  $x$  in multiset  $Y$  i.e.  $n(y, Y^u)$  takes duplicates into account.  $N'_0 = \sum_{y \in Y_0^u} n(y, Y_0^u)$ ,  $N'_1 = \sum_{y \in Y_1^u} n(y, Y_1^u)$  are normalising constants,  $\rho(y, Y^u) = \max_{x \in N_k(y, Y^u)} \|x - y\|$  is the distance to the  $k$ 'th neighbor and  $d$  is the dimension of the vector embeddings. This  $k$ NN approach extends the continuous pdf estimator of [4] to include probability distributions with both discrete and continuous parts. The estimator in [4] is known to be consistent and to scale well to high dimensional data (embedding are typically have dimension  $d$  around 1000).

## 4.2 Estimating Renyi-Divergence

To estimate the Renyi-divergence between two datasets we extend the estimator of [4], which is observed to scale well for high dimensional data. We updated

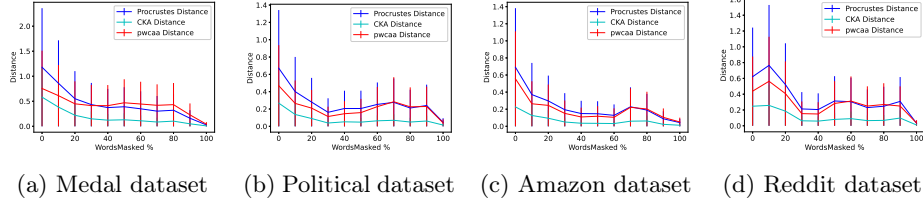


Fig. 5: Measured CKA,PWCAA and Procrustes distances between the decoder weights of the model trained on sensitive and safe dataset. A lower value indicates the weights are similar.

the estimator to handle duplicate word-sequences, since these can become common following redaction. In addition, each word sequence is mapped to a vector embedding to work well. See Section-[Choice of Embedding form the paper]  $X_i$ . 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.

### 4.3 Impact of Redaction on Renyi-Divergence

Applying redaction policy  $\pi_p$  to both the sensitive dataset  $\mathcal{D}_1$  and the safe dataset  $\mathcal{D}_0$  then we expect that distance  $\epsilon$  between the datasets decreases as the level of redaction increases, the distance becoming zero at redaction  $p = 1$  (in which case the elements in both datasets degenerate to be the same uninformative mask token MASK).

Figure 2 illustrates this behaviour 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. The figure shows the measured Renyi-divergence  $D_2(\hat{P}_0||\hat{P}_1)$  between the empirical probability distributions  $\hat{P}_0$  and  $\hat{P}_1$  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.

### 4.4 Calculating $(\epsilon, \delta)$

To calculate  $\xi$  and  $\rho$  in equation - 2 (from the paper), we first calculate  $D_\alpha$  for a range of  $\alpha$  values<sup>2</sup>. We then find a line that lies above the  $D_\alpha$  vs  $\alpha$  curve and

<sup>2</sup> We select the range to be large enough that  $D_\alpha$  no longer increases as we increase  $\alpha$ .

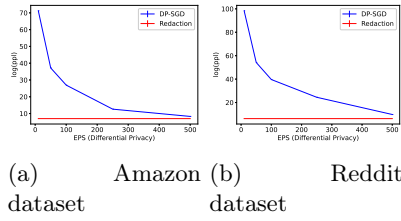


Fig. 6: Measured impact of privacy on utility. Next word prediction performance for LSTM trained on redacted dataset and using DP-SGD. x axis represents  $\epsilon$  value of privacy guarantee and y axis represents  $\log(\text{perplexity})$

select  $\rho$  as the slope of the line and  $\xi$  the intercept. Of course, many lines lie above the  $D_\alpha$  curve, so we try to select one such that  $\rho$  and  $\xi$  are minimised. See for example Figure-3, which shows  $D_\alpha$  vs  $\alpha$  for the Medal medical dataset (the blue curve). This curve is upper bounded by the red line. The values of  $\rho$  and  $\xi$  corresponding to this red line are plugged into equation-3 (from the paper) to obtain the corresponding  $(\epsilon, \delta)$  privacy values.

## 5 KL Divergence

The Renyi-divergence estimator can be extended to calculate KL-divergence between the sensitive and safe datasets. Figure-1, plots the measured KL divergence between the sensitive and safe datasets as the redaction level is increased using smarter redaction.

## 6 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 embedding back to a word [2, 6, 1]. 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<sup>3</sup>.

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

<sup>3</sup> 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 [3].

Figure 4 shows the measured  $\epsilon$ -renyi for the Medal dataset as the level of noise is increased (indicated by  $\epsilon$ ) 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  $\epsilon$  values), both the  $\epsilon$ -renyi values and the attack accuracy remain high.

## 7 Closeness of Model

Figure-5 plots for the distances between weights of the decoder of the models trained on sensitive and safe dataset. We can see that both the encoder and decoder weights of the model comes closer as we increase redaction.

## 8 Utility Over Wider $\epsilon$ Range

Figure-6 shows the measured perplexity of the DP-SGD and redaction over a wider range of  $\epsilon$  value than shown in Figure-7d (which focuses on the low  $\epsilon$  regime of most interest for privacy). It can be seen that, as expected, the performance of DP-SGD and redaction eventually become the same for high enough  $\epsilon$ , and the performance at this point matches the performance when the model is trained without privacy.

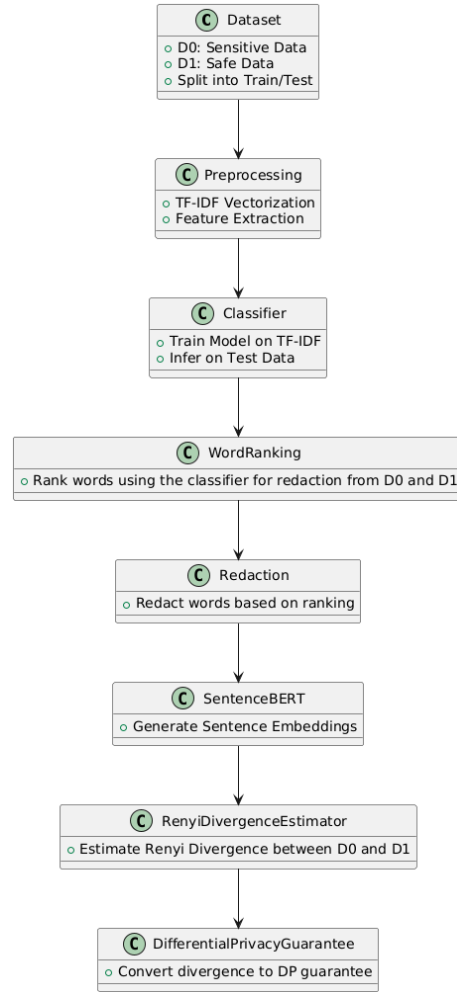


Fig. 7: An overview of the redaction pipeline.

## Bibliography

- [1] Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [2] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20*, page 178–186, New York, NY, USA, 2020. Association for Computing Machinery.
- [3] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD '11*, page 193–204, New York, NY, USA, 2011. Association for Computing Machinery.
- [4] Morteza Noshad, Kevin R. Moon, Salimeh Yasaei Sekeh, and Alfred O. Hero. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, June 2017.
- [5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- [6] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online, August 2021. Association for Computational Linguistics.