Text Redaction for Privacy Preserving Model Training- Appendix.

Vaibhav Gusain[1] and Douglas Leith[1]

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 $C = 1.5$.

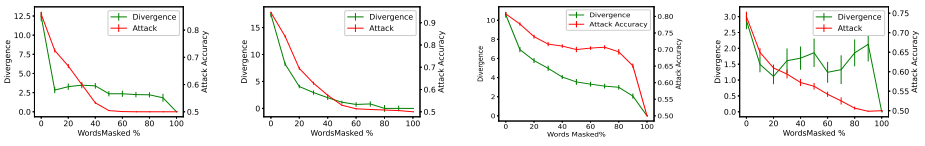
3 Overview of the redaction pipeline

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 α between two probability distributions P_0 and P_1 on sample space Y is given by) equation 1 (in the paper). We need to accommodate probability distributions with both discrete and continuous parts since



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

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.

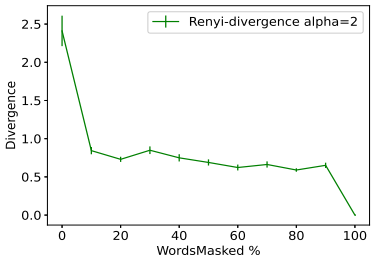


Fig. 2: Measured Renyi divergence for ($\alpha = 2$) vs redaction level for Medal dataset.

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 [1] 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 multset 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

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

[1] 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 (**?**)

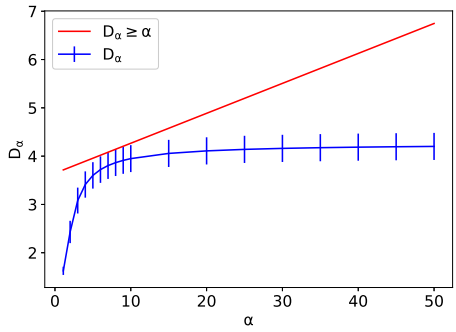
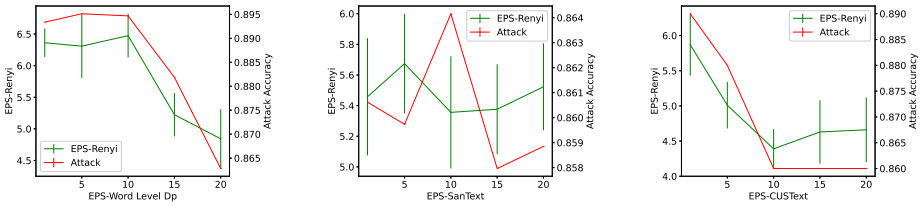


Fig. 3: Divergence vs α for non-redacted cancer and non-cancer text from Medal medical dataset.



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

Fig. 4: Measured ϵ-renyi and attack accuracy for various word-level DP approaches applied to Medal Dataset.

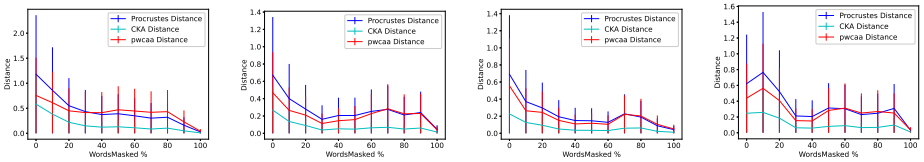
with

$$\hat{P}_0(y) = \frac{n(y, Y_0^u)}{N_0' \rho(y, Y_0^u)^d}, \hat{P}_1(y) = \frac{n(y, Y_1^u)}{N_1' \rho(y, Y_1^u)^d}$$

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 kNN approach extends the continuous pdf estimator of (**?**) to include probability distributions with both discrete and continuous parts. The estimator in (**?**) 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 **?**, which is observed to scale well for high dimensional data. We updated



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

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

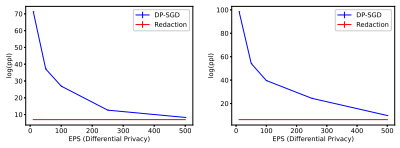
Applying redaction policy π_p to both the sensitive dataset $\mathcal{D}_1$ and the safe dataset $\mathcal{D}_0$ then we expect that distance ϵ between the datasets decreases as the level of redaction increases, the distance becoming zero at redaction $p = 1$ (in which case the elements in both datasets 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 $D_2(\tilde{P}_0||\tilde{P}_1)$ between the empirical probability distributions $\tilde{P}_0$ and $\tilde{P}_1$ 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 (ϵ,δ)

To calculate ξ and ρ in equation - 2 (from the paper), we first calculate D_α for a range of α values[2]. We then find a line that lies above the D_α vs α curve and

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



(a) Amazon dataset (b) Reddit dataset

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 ϵ value of privacy guarantee and y axis represents $log(perplexity)$

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 (from the paper) to obtain the corresponding (ϵ,δ) 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 (**???**). 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[3].

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.

[3] 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 (**?**).

Figure 4 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.

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 ϵ Range

Figure-6 shows the measured perplexity of the DP-SGD and redaction over a wider range of ϵ value than shown in Figure-7d (which focuses on the low ϵ 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 ϵ, and the performance at this point matches the performance when the model is trained without privacy.

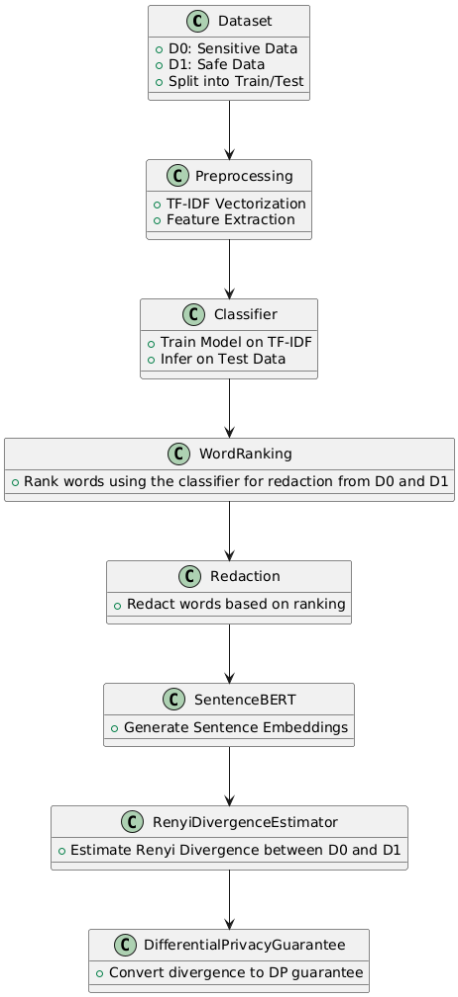


Fig. 7: An overview of the redaction pipeline.