Improving Privacy Benefits of Redaction-Appendix

Vaibhav Gusain[1] and Douglas Leith[1]

1- Trinity College Dublin - School of Computer science and statistics
Dublin - Ireland

1 Privacy Defination (ϵ, δ)

Considering two datasets $\mathcal{D}_0$ and $\mathcal{D}_1$. 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 $\mathit{redact}(x)$ and the redacted dataset becomes a sample from probability distribution $\mathit{redact}(P)$. The distance between two redacted datasets $\mathit{redact}(\mathcal{D}_0)$ and $\mathit{redact}(\mathcal{D}_1)$ can be measured 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 $\tilde{P}_0 := \mathit{redact}(P_0)$ is the probability distribution over token sequences in dataset $\mathit{redact}(\mathcal{D}_0)$, $\tilde{P}_1 = \mathit{redact}(P_1)$ in dataset $\mathit{redact}(\mathcal{D}_1)$ 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 $\mathit{redact}(\mathcal{D}_0)$ rather than all databases differing from $\mathit{redact}(\mathcal{D}_1)$ by a single element. When ϵ, δ are sufficiently small, the publication of private dataset $\mathit{redact}(\mathcal{D}_1)$ 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 $\mathit{redact}(\mathcal{D}_0)$ and $\mathit{redact}(\mathcal{D}_1)$ 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 concentrated differential privacy. See [1] for more details.

2 Outline for algorithm

Algorithm-1 shows a pseudocode for training the ranker.

3 Hyper parameters

The ranker was trained for 3 epochs. Batch size of 64 sentences was used during training. Adam optimizer was used with a learning rate of 0.001. Ranker model consists of 4 linear layer, the first three layers having tanh activation and the final layer having a sigmoid activation function.

The Full training code is available here - `add github repo`

Algorithm 1 Train Ranker. D_0 represents the sensitive dataset. D_1 represents the safe dataset. sent_trans is the sentence transformer fine-tuned on the training data. lossfn is the custom KL-divergence loss explained in Section-**??**. bsz is the batchsize. sgd is stochastic gradient descent algorithm which is used to update the weights.

```
function : train_ranker($D_0, D_1, sent\_trans, model, lossfn, bsz, sgd$)
    $k \leftarrow 0$ ; $N \leftarrow len(D_0)$
    $randomize(D_0)$; $randomize(D_1)$
    while $k \leq N$ do
        $db_0 \leftarrow D_0[k : k + bsz]$ ; $db_1 \leftarrow D_1[k : k + bsz]$
        $e_0 \leftarrow sent\_trans(db_0)$ ; $e_1 \leftarrow sent\_trans(db_1)$
        $r_0 \leftarrow model(e_0)$ ; $r_1 \leftarrow model(e_1)$
        $ue_0 \leftarrow e_0 \cdot r_0$ ; $ue_1 \leftarrow e_1 \cdot r_1$
        $loss \leftarrow lossfn(ue_0, ue_1)$
        $model \leftarrow sgd(model, loss)$
        $k \leftarrow k + bsz$
    end while
    return $model$
end function
```

4 Complete information about the datasets

In this section we provide a complete overview about the datasets used for experiments :

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

(i) Medal dataset [3][1]. 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- [2][2]. This contains comments on Facebook posts from 412 members of the United States Senate and House. Each comment is labeled with the corresponding Congresspersonâs 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[3]. This dataset contains product reviews from Amazon customers. We selected the reviews which were categorised as ”drug-store” and

[1] `https://huggingface.co/datasets/medal`

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

[3] `https://huggingface.co/datasets/amazon_reviews_multi`

"kitchen-appliances". For our experiments, the dataset with drug-store reviews was chosen to be the sensitive dataset.

(iv) Reddit dataset[4]. 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.

References

[1] Vaibhav Gusain Douglas Leith. Plausible deniability of redacted text. In *DPM International Workshop on Data Privacy Management, ESORICS 2024*, 2024.

[2] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[3] Zhi Wen, Xing Han Lu, and Siva Reddy. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 2020.

[4] `https://www.kaggle.com/general/256134`