Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Plausible Deniability of Redacted Text

Vaibhav Gusain, Douglas Leith

Date 19/09/24

Outline

- Introduction
- Related work
- Experiments
- Results
- Discussions

Introduction

- **Problem Statement:**
  - How can we sanitize input text data so that it can be used for model training while preserving privacy
- **Current approaches**
  - DP-SGD
  - Sanitise Input data
    - For textual dataset : The input word is mapped to an embedding vector. Noise is added to the vector and the noisy embedding is then mapped back to the original word.
- **Sanitizing Input text is difficult.**
  - Choice of embedding greatly affect the amount of noise added and thus the final privacy
  - We can’t simply replace the words as the surrounding context can reveal sensitive information
    - “I am diagnosed with cancer. I have to go to St Lukes for chemotherapy and will probably lose my hair”.
- **Our approach**
  - Instead of focussing on individual words or sentences, we worked on the whole text corpus.
  - Use redaction to add “noise” to the text.

Plausible Deniability of Redacted Text

Privacy Metric - Renyi Divergence

- We used embeddings from a sentence transformer
- Assuming a “safe” dataset and sensitive dataset, redacting words from the sensitive and safe dataset reduces the divergence between the two datasets.
- Selecting the proper level of redaction to ensure sufficiently small divergence then provides privacy in the sense of indistinguishability between the redacted sensitive and safe dataset.

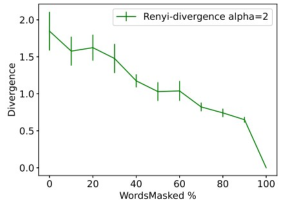
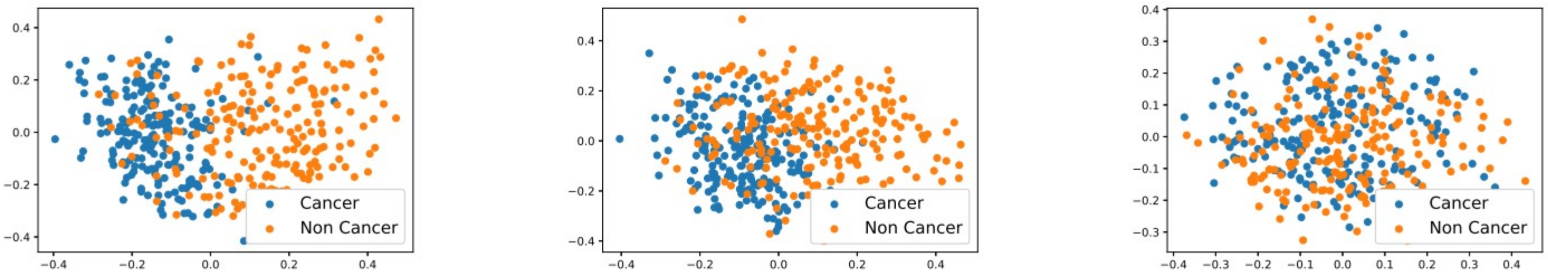


Fig. 1: Measured Renyi vs random redaction level for Medal dataset.

Plausible Deniability of Redacted Text



(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.

Plausible Deniability of Redacted Text

- **Threat model**
  - Attacker has access to the redacted datasets and wants to infer the sensitive traits from them.
- **Redaction**
  - Random - randomly redact X% of the words from the input sentence
  - Smarter redaction - Use a logistic regression classifier to weigh important words and mask important words first.

Renyi-Divergence, Zero Concentrated Differential Privacy

Renyi Divergence of order **α** between two probability distributions P_0 and P_1 is

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

We say that the probability distributions P_0 and P_1 are (ξ,ρ)-zero-concentrated differentially private when :
For all **α** ∈ (1,∞).

$$D_\alpha(P_0||P_1) \leq \xi + \rho\alpha$$

Probability distributions P_0 and P_1 are (ξ,ρ)-zero-concentrated differentially private then:

$$P_1(x) \leq \exp(\epsilon) P_0(x) + \delta$$

Where for every **δ** > 0, $\epsilon = \xi + \rho + 2\sqrt{\rho \log \frac{1}{\delta}}$.

*Bun, M., Steinke, T.: Concentrated differential privacy: Sir
extensions, and lower bounds (2016)

Divergence to epsilon

- For a given redacted safe and sensitive dataset. Calculate renyi divergence for different alphas.
- Plot the curve of divergence v alpha.
- Get a line which is above the plotted curve
  - ρ is the slope of the line and ξ is the intercept.
- Using Zero-concentrated-differential privacy calculate the (ε,δ) differential privacy guarantee.

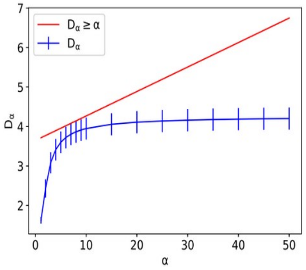
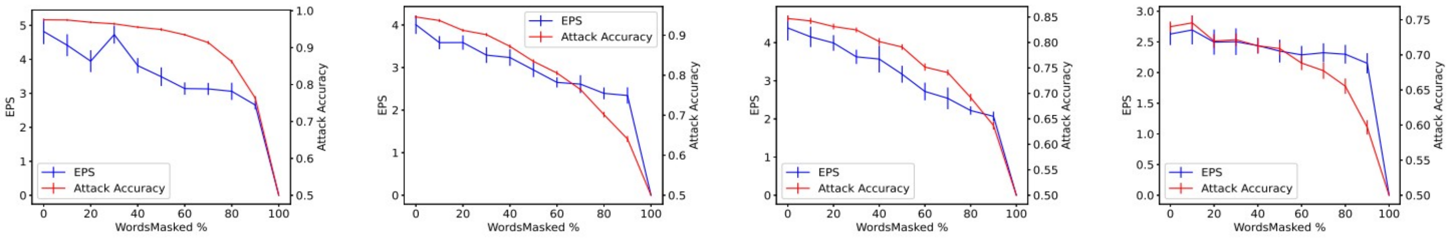


Fig. 3: Divergence vs α for non-redacted cancer and non-cancer text from Medal medical dataset.

Plausible Deniability of Redacted Text

Redaction and Attack accuracy

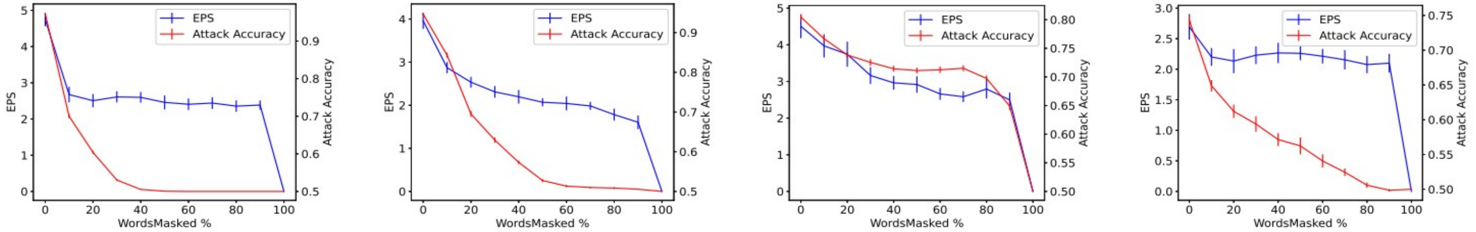


(a) Medal dataset (b) Political dataset (c) Amazon dataset (d) Reddit dataset

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).

Plausible Deniability of Redacted Text

Redaction and Attack accuracy

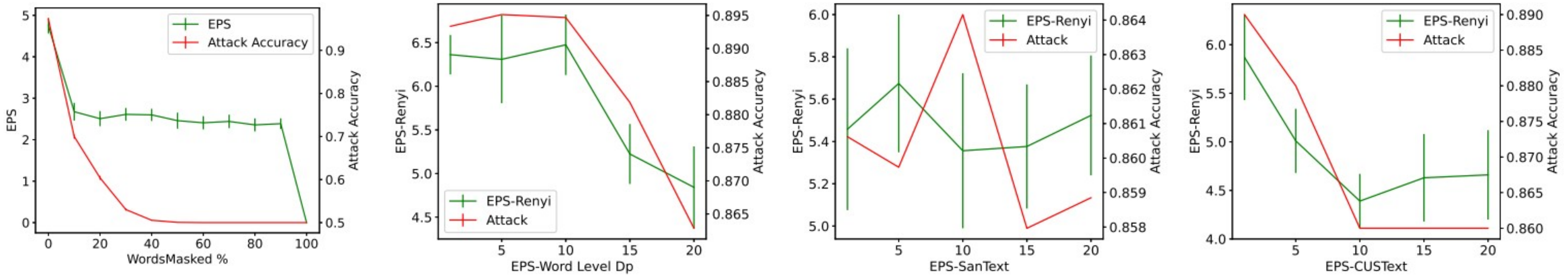


(a) Medal dataset (b) Political dataset (c) Amazon dataset (d) Reddit dataset

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).

Plausible Deniability of Redacted Text

Comparison against State of the Art



(a) Our Approach (b) Word Level DP (c) Santext plus (d) Custext

Fig. 6: Comparison against various SOTA approaches.

Discussions

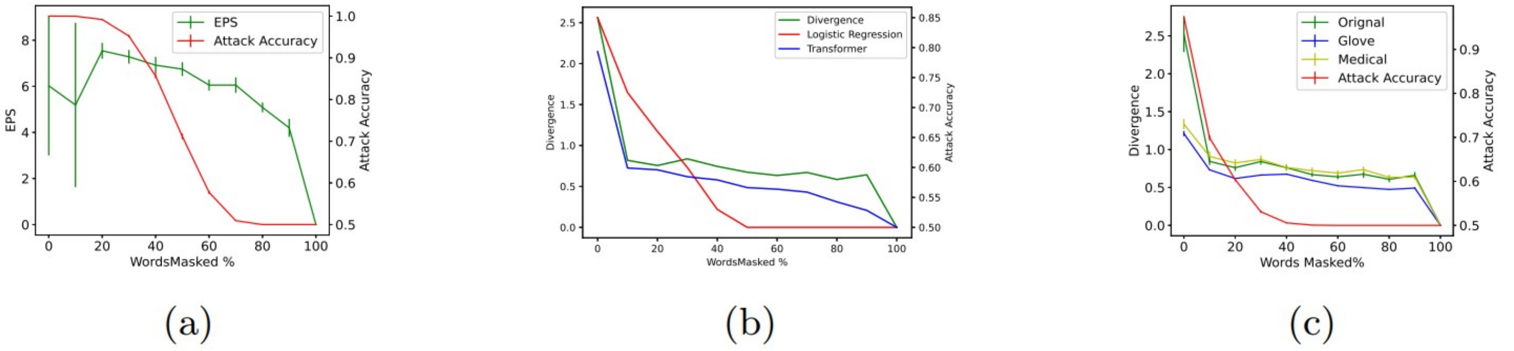


Fig. 8: 8a Measured ϵ and attack accuracy for cancer sentences when compared against IMDB reviews.8b Measured Renyi-divergence ($\alpha = 2$) and attack accuracy for logistic regression and BERT transformer classification attacks as the redaction level is increased. Medal dataset. 8c Measured Renyi-divergence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

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You