BRR: Preserving Privacy of Text Data Efficiently on Device

Ricardo Silva Carvalho[1], Theodore Vasiloudis[2], Oluwaseyi Feyisetan[2]

[1]Simon Fraser University [2]Amazon

Summary

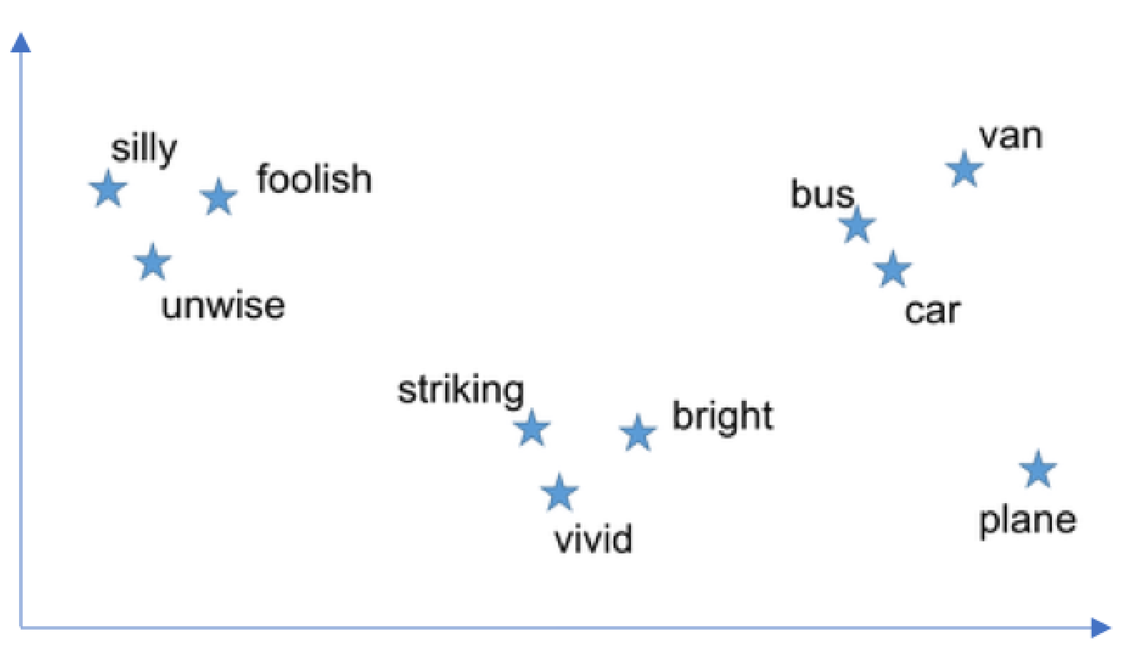
We propose an efficient mechanism to provide metric differential privacy for text data on-device. Our contributions are the following:

- New zero-trust algorithm for on-device text privatization, using binary embeddings and randomized response.
- Theoretical method to compare metric DP mechanisms that use different metrics.
- Empirical evaluation demonstrating the computational advantages of the approach compared to the state-of-the-art, while maintaining better or similar utility.

Introduction

Metric Differential Privacy: Framework to give formal privacy guarantees generalized to use with a metric space.

Privatizing Words: Ensuring the privacy of users whose data are used to train Natural Language Processing (NLP) models. Usually representing words via embedding vectors.



Metric Differential Privacy

Given a distance metric $d : \mathcal{W} \times \mathcal{W} \to \mathbb{R}_+$, a randomized mechanism $\mathcal{M} : \mathcal{W} \to \mathcal{Y}$ is εd-DP if for any $w, w' \in \mathcal{W}$ and all outputs $y \in \mathcal{Y}$:

$$\Pr[\mathcal{M}(w) = y] \leq e^{\varepsilon d(w,w')} \Pr[\mathcal{M}(w') = y]$$

Existing method: Madlib

The previous state of the art algorithm, Madlib [1], had the following characteristics:

- Privatization was done by adding noise to inputs in the metric space of word embeddings.
- Uses a continuous distance metric: Euclidean.
- Assumes a trusted central authority that gathers the data of all the users.

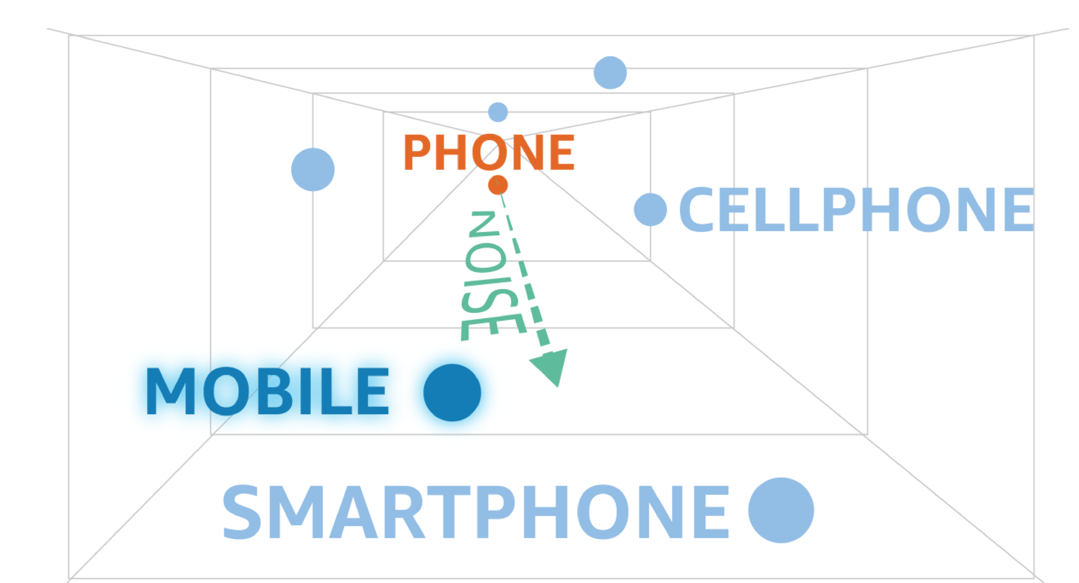


Figure: Madlib adds noise to the inputs's embedding vector and finds nearest neighbor.

Our method: BRR

Our method, BRR, satisfies the following:

- Uses **binary** embedding vectors to represent words and applies Randomized Response (RR) to make each vector differentially private.
- Binary vector representations follow [2].
- Assumes zero-trust, i.e. does not need any trusted party to gather sensitive data.

Algorithm 1 - BRR: Mechanism for Text as Binary Embeddings over Randomized Response

Input: Finite domain $\mathcal{W}$, input word $w \in \mathcal{W}$ and privacy parameter ε.

Output: Privatized word $\hat{w}$.

1: Compute **binary** embedding vector $\phi_w = \phi(w)$
2: Perturb word embedding vector using **Randomized Response** to obtain $\hat{\phi}_w = RR(\phi_w, \varepsilon)$
3: Obtain perturbed word:
4: $\hat{w} = \text{argmin}_{y \in \mathcal{W}} \left\| \phi(y) - \hat{\phi}_w \right\|$
5: Return $\hat{w}$

Comparing mDP Mechanisms

- We have that for any $x, x' \in \mathcal{X}$ and all outputs $y \in \mathcal{Y}$: $\mathcal{L}_{\mathcal{M},x,x'}(y) < \varepsilon \cdot d(x,x')$
- Thus we estimate the *privacy loss bound* via $\varepsilon \cdot \mathcal{P}_d$, where $\mathcal{P}_d$ is defined as an aggregate distance measurement based on the distances between all possible pairs of words.
- To fairly compare the privacy of two mechanisms, we equalize their bounds via a privacy ratio.
- For any given ε_A defined for $\mathcal{M}_A$ we need to set: $\varepsilon_B = \mathcal{R}_{d_A,d_B} \cdot \varepsilon_A$ where $\mathcal{R}_{d_A,d_B} = \mathcal{P}_{d_A}/\mathcal{P}_{d_B}$.

Utility Results

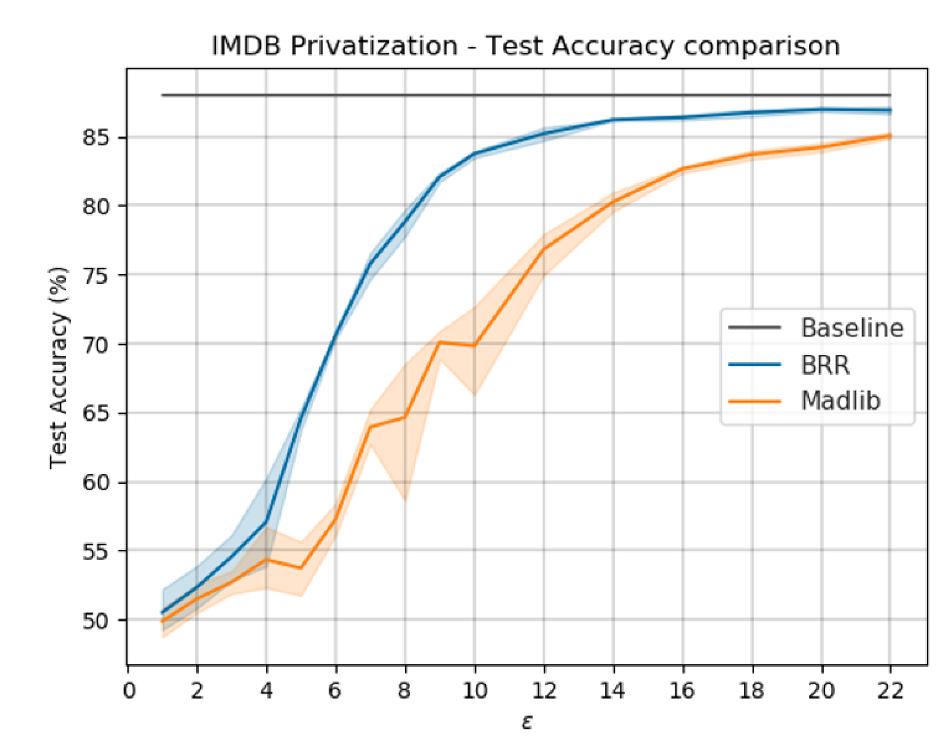


Figure: Test accuracy of sentiment analysis models trained on privatized data. Baseline is model trained on sensitve data.

Privacy Results

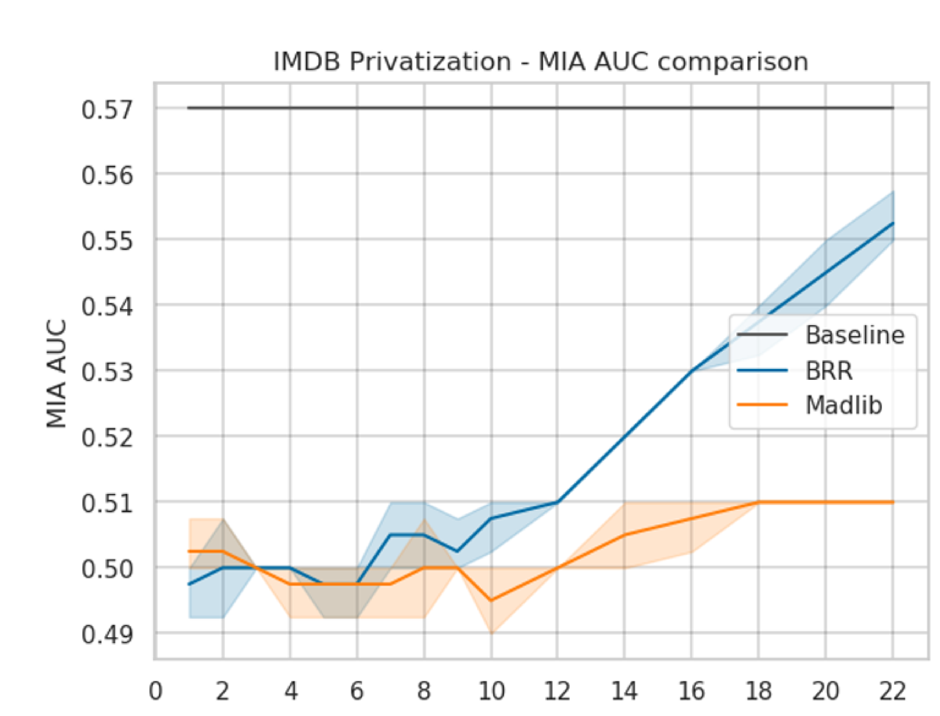


Figure: AUC of Membership Inference Attacks on models. Smaller is better. Shows that both mechanisms preserve privacy.

Efficiency Results

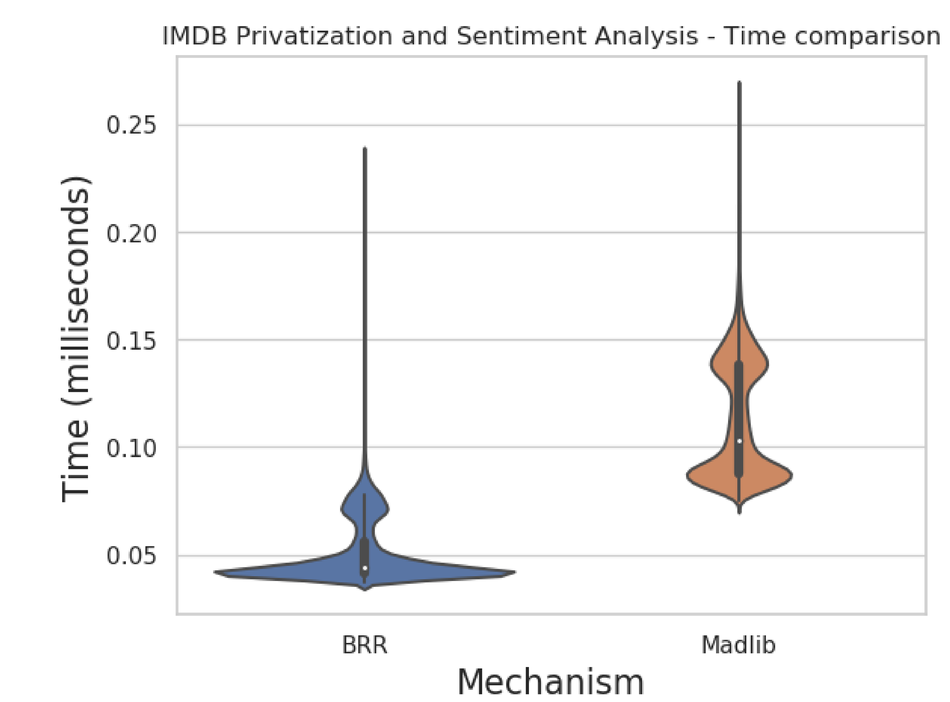


Figure: Wall-time: BRR vs Madlib on IMDB dataset.

- In our experiments, BRR was on average 68% faster to privatize a word compared to Madlib.
- NN index: Compressed 97.9% (4MB vs. 200MB).
- Vocab. + embeddings: 98.5% (6MB vs. 300MB).

Conclusion

- We presented an efficient zero-trust mechanism for text privatization on-device with formal differential privacy guarantees.

References

[1] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. WSDM, 2020.

[2] Julien Tissier, Christophe Gravier, and Amaury Habrard. Near-lossless binarization of word embeddings. AAAI, 2019.

Acknowledgments

This work was performed while Ricardo was completing an internship at Amazon Web Services.

Contact

- rsilvaca@sfu.ca
- thvasilo@amazon.com
- sey@amazon.com

SFU SIMON FRASER UNIVERSITY amazon.com