TruncatedEM: High Utility Metric Differential Privacy on Text

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

[1]Simon Fraser University [2]Amazon

Summary

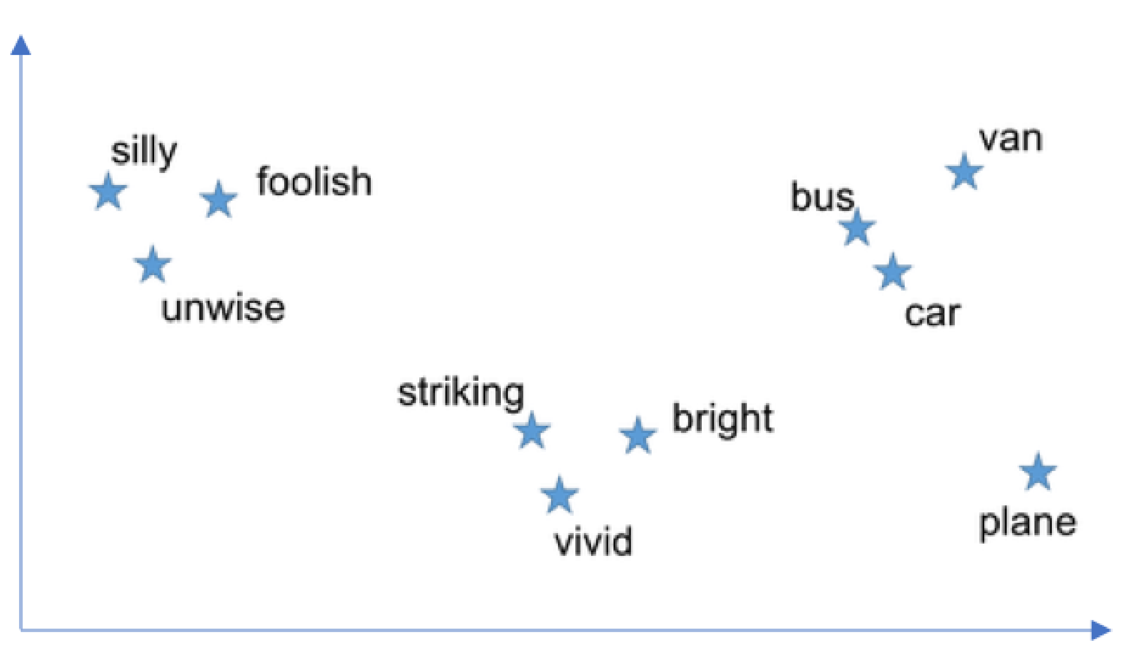
We propose a method satisfying metric differential privacy for word privatization using any distance metric on sensitive word embeddings.
Our contributions are the following:

- New method that adjusts noise to regions on the embedding space for better utility.
- Added truncation step to initially select from high utility words with tunable error.
- Allows pre-processing for computationally efficient word selection.

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.
- Assumed the embeddings used have been trained on non-sensitive data.
- Only worked with a pre-defined distance metric.
- Ignored the density of the space around the input.

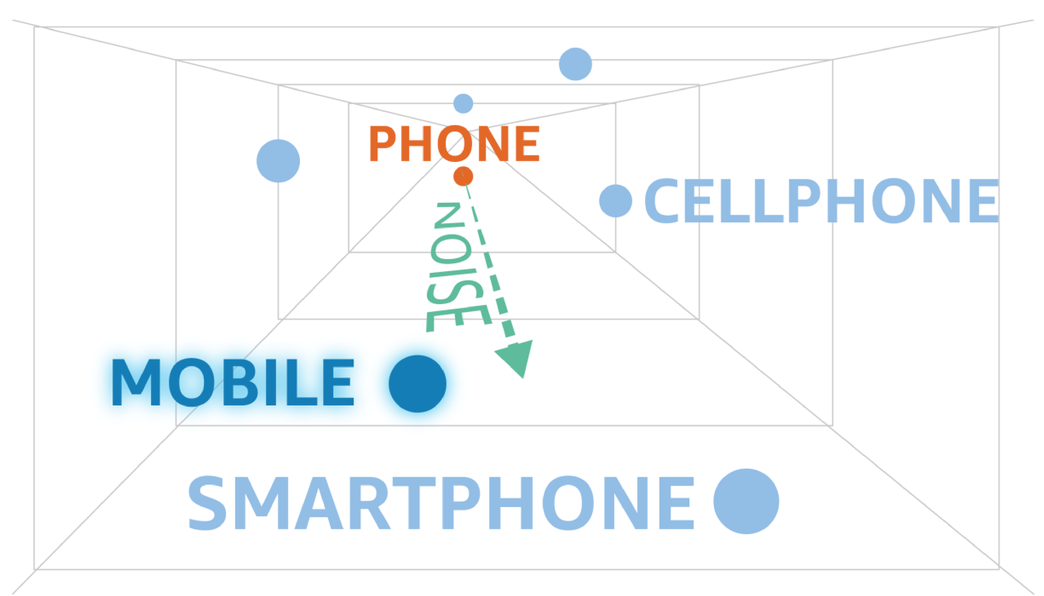


Figure: Madlib poses word privatization as a vector release problem.

Our method: TruncatedEM

Our method, TruncatedEM, selects words from within a radius from the input word, giving higher score to words closer to the input.

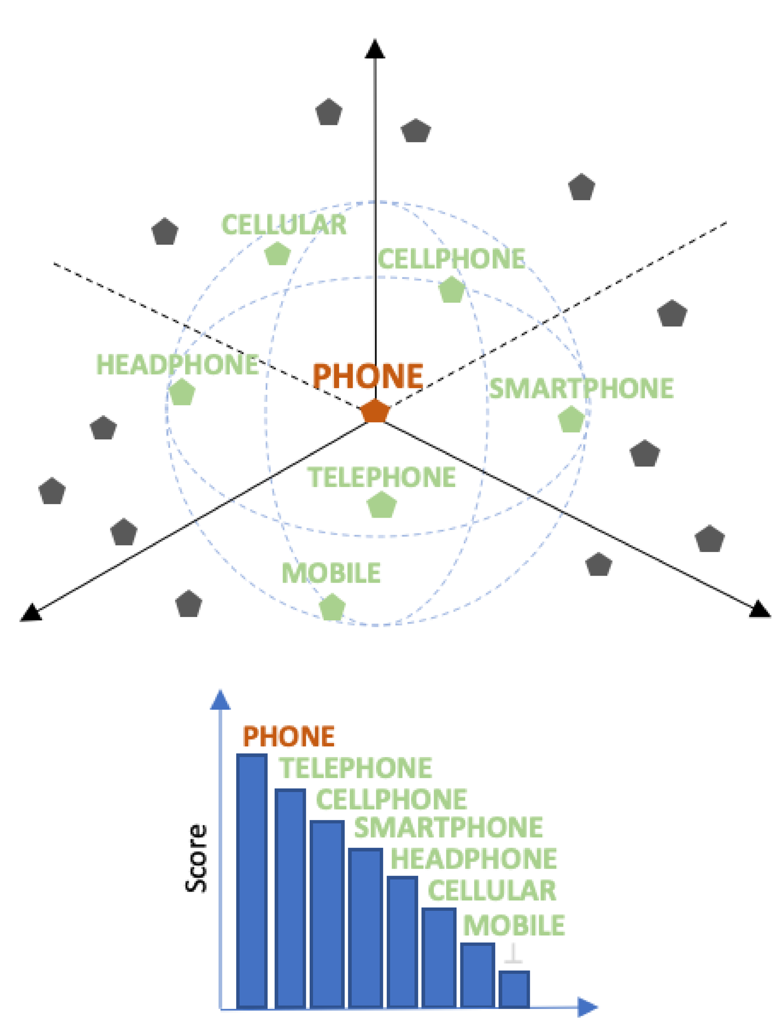


Figure: TruncatedEM poses word privatization as a selection problem.

TruncatedEM versus Madlib

- Dynamic noise behavior, adapted to density.
- Allows embeddings trained on sensitive data.
- Works with any formal distance metric.
- Includes an optional pre-processing step for improved computational efficiency.

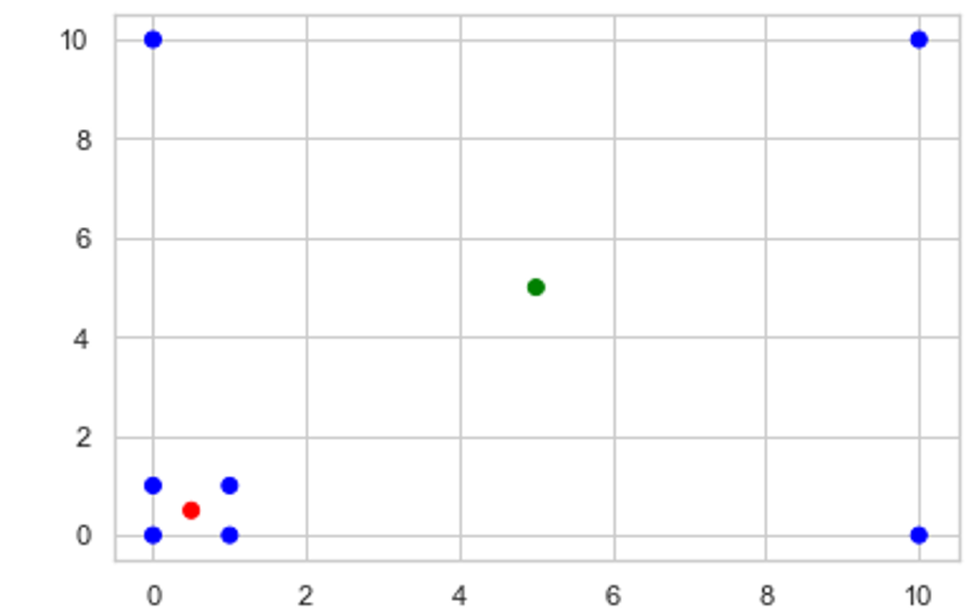


Figure: TruncatedEM adds less noise to high density areas (red dot) and more noise to low density areas (green dot).

Threshold for TruncatedEM

Setting γ based on error parameter: For error probability $\beta > 0$, $w \in \mathcal{W}$, TruncatedEM outputs elements with distance at most γ from input w with probability at least $1 - \beta$ for:

$$\gamma \geq \frac{2}{\varepsilon} \cdot \ln \frac{(1-\beta)(|\mathcal{W}|-1)}{\beta}$$

Utility Results

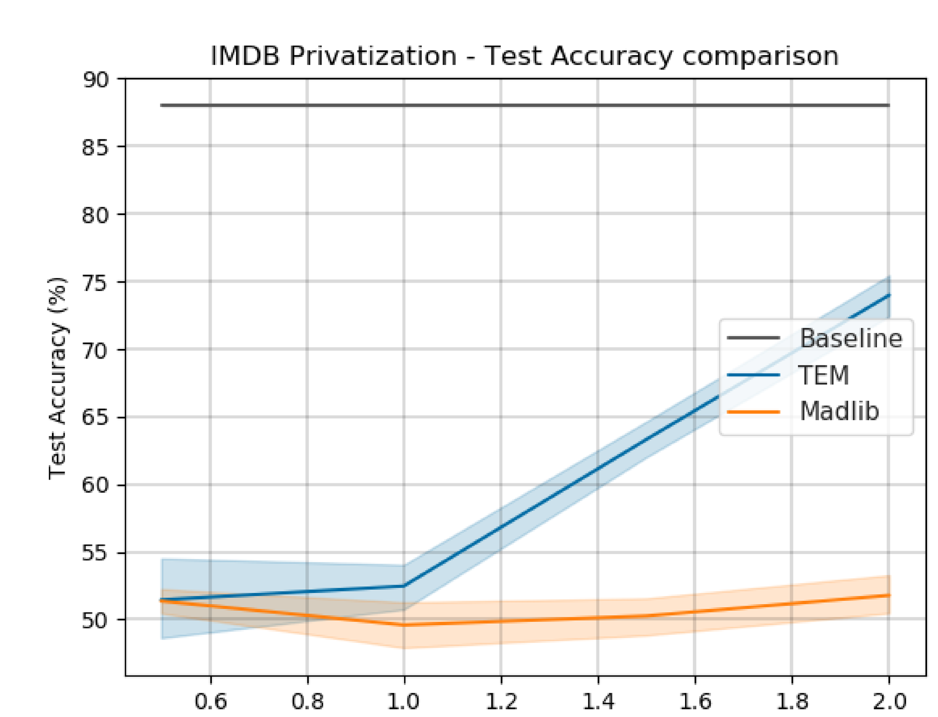


Figure: Test accuracy of sentiment analysis models trained on privatized data. Baseline is model trained on sensitve data.

Privacy Results

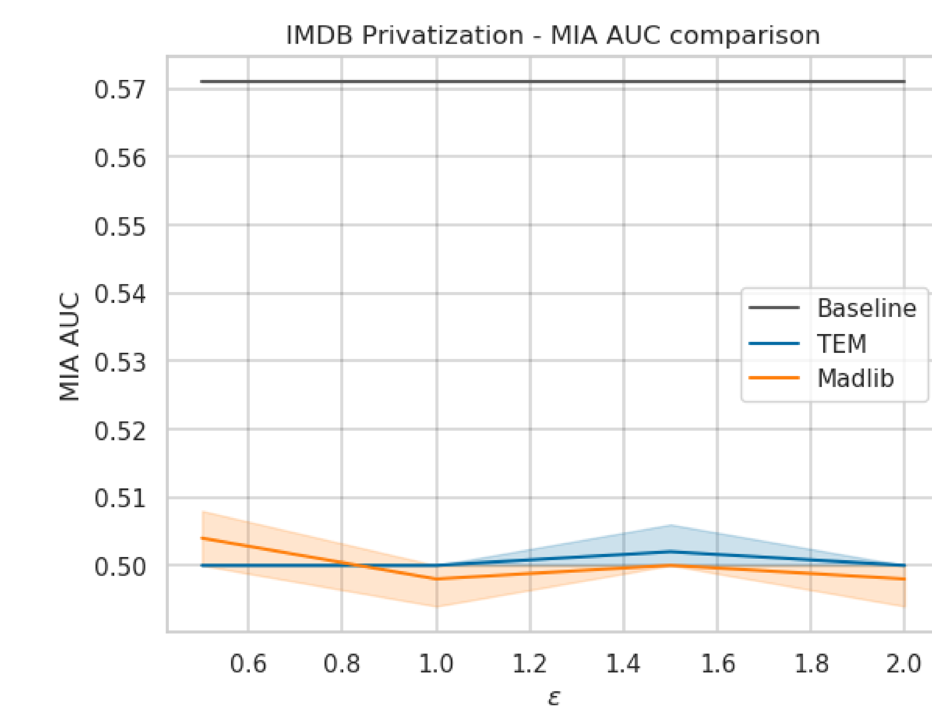


Figure: AUC of Membership Inference Attacks on models. Smaller is better. Shows that both mechanisms preserve privacy.

Conclusion

- In this work we proposed an efficient, high utility, text privatization mechanism for any distance metric with adaptive noise that allows the use of sensitive embeddings.

References

[1] 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.

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