



CORIOLIS
TECHNOLOGIES

Industry Project Internship Report

Authored By:
Divyanshi Kumari
MDS202322

Acknowledgement

I would like to sincerely thank **Mr. Besant Rajan, CEO of Coriolis Technologies**, for granting me the opportunity to work on this industry project.

I am deeply grateful to **Mr. Sudhir Kumar** and **Mr. Rohan Nandode** for their constant mentorship and invaluable guidance throughout this journey—especially during moments of challenge and saturation.

I would like to express my gratitude to **Prof. Pranabendu Misra** for taking time out of his busy schedule to assist us in exploring advanced techniques and deepening our understanding of the subject.

My heartfelt thanks also go to the entire Coriolis team for their insightful feedback and for helping evaluate the outcomes of the project.

This internship has been an enriching experience that allowed me to broaden my knowledge, explore new concepts, and apply my learning to real-world challenges.

Index

S.No	Topic	Page
1	Abstract	4
2	Introduction	4
3	Background	5
4	Objectives	5
5	Data	5
6	Methodologies	6
7	Experimental Setup	6
8	Results & Discussions	7
9	DP-Forward	9
10	Conclusion	10
11	Appendix	10

1. Abstract

Pretrained language models like BERT encode semantic information into dense embeddings. Privacy risks in such pretrained language models have raised concerns about sensitive information leakage through embeddings. In this project, we investigate lightweight, post-facto defenses that perturb embeddings to enhance privacy while preserving utility. Our techniques — including Gaussian noise addition, random projection, and random rotation — achieved up to a **50% reduction** in embedding inversion attack success while maintaining cosine similarity within a **10% deviation** threshold. Crucially, our methods operate externally without modifying BERT's architecture, enabling seamless deployment.

2. Introduction

Pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized natural language processing (NLP) by learning rich semantic representations from large text corpora. BERT generates embeddings that capture the meaning and structure of text, making them highly valuable for a wide range of downstream applications.

BERT embeddings are widely used in:

- Search engines (semantic search, ranking)
- Conversational AI (chatbots, virtual assistants)
- Healthcare and legal document analysis
- Recommendation systems

Their effectiveness and generalization make embeddings a common component in many production systems and APIs.

However, this creates a **privacy risk**. Through an **Embedding Inversion Attack**, an adversary can attempt to reconstruct the original input text from its embedding. By training a decoder or using optimization techniques, attackers can reveal sensitive information encoded in the embeddings.

In high-stakes domains like healthcare or finance, even partial reconstruction can lead to serious data leaks. Therefore, protecting embeddings against inversion attacks — while maintaining their utility — is critical. In this work, we explore lightweight perturbation techniques inspired by Differential Privacy to defend against such attacks.

3. Background

Differential Privacy

A randomized mechanism M provides (ϵ, δ) -differential privacy if for any two neighboring databases, D_1 and D_2 , that differ in only a single record, and for all possible outputs $S \subseteq \text{Range}(A)$:

$$P[M(D_1 \in A)] \leq e^\epsilon P[M(D_2 \in A)] + \delta$$

ϵ : the privacy budget, is a metric of privacy loss. lower ϵ values indicate higher levels of privacy but are likely to reduce utility as well.

δ : accounts for a small probability on which the upper bound ϵ does not hold.

In the simplest setting, consider an algorithm that analyzes a dataset and computes statistics about it (such as the data's mean, variance, median, mode, etc.). Such an algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not.

The major challenge that Differential Privacy introduces is the tradeoff between utility and privacy, a balance we have carefully navigated throughout the project, with our primary goal being to preserve the usefulness of the embeddings while also enhancing their protection.

4. Objectives

This project explores the application of Differential Privacy (DP) inspired techniques to BERT embeddings. The primary objectives are two-fold:

1. **Utility Preservation:** Maintain cosine similarity between sentence embeddings within a 10% deviation of the original.
2. **Privacy Protection:** Prevent inference attacks—especially embedding inversion attack

5. Data

In this project, we primarily work with a subset of the Semantic Textual Similarity Benchmark (STS-B) dataset, using 100 sentence pairs from the test set for initial experiments. To further validate the effectiveness and generalizability of our privacy-preserving techniques, we also extend our evaluation to larger datasets: 500 sentence pairs from the STS-B validation set and 500 pairs from the Quora Question Pairs (QQP) test set. These additional evaluations help ensure that our methods are not overfitted to a small sample and maintain consistent performance across different datasets.

6. Methodologies

1. Post-Facto Noise Addition

- Extract embeddings from a pretrained BERT model and perturb them with random noise
- **Noise types:** Standard Gaussian, orthogonal Gaussian, Laplace, and uniform noise. $z' = z + n, n \sim \text{NoiseDistribution}$
- Optimal noise scales were chosen empirically to satisfy cosine similarity constraint.
- **Evaluation:** Test robustness of noisy embeddings using embedding inversion attacks.

2. Random Projection

- Multiply the embedding by a random Gaussian matrix: $z' = Pz$ where $P \sim N(0, 1)$
- Approximate distance preservation due to Johnson–Lindenstrauss Lemma.

3. Random Rotation

We apply random pairwise 2D rotations to embeddings, using per-sample random angles, to perturb embeddings while preserving local geometric structure.

4. Squaring(certain embedding dimensions)

Certain embedding dimensions are squared with sign preservation to introduce controlled non-linear perturbations

$$z'_i = \text{sign}(z_i) \times z_i^2$$

7. Experimental Setup

7.1 Datasets

- **STS-B Test Set:** 100 sentence pairs
- **STS-B Validation Set:** 500 pairs
- **Quora Question Pairs Test Set:** 500 pairs

7.2 Attack Script

- **Embedding Inversion Attack:** mapping to reconstruct original embeddings from noisy versions.
- Attack trained for **1 epoch** for evaluation consistency.
- Increased training results are in the Experiments summary in Appendix

7.2 Metrics

- **Cosine Similarity Deviation:** Between original and perturbed embeddings.
- **Token Retrieval Rate:** Number of correctly retrieved tokens during attack.
- **% Reduction in Attack Success.**

8. Results and Discussion

Post-Facto Noise

Noise Type	Optimal Noise_Scale	Tokens Retrieved	Total Tokens	Deviation in cosine-similarity	% reduction in attack success
Unprotected BERT (Baseline)	-	86	407	-	-
Gaussian (std)	0.07	73	407	6.85	15%
Gaussian(orthogonal projection)	0.3	78	407	7.0	9%
Laplace	0.05	77	407	6.8	10%
Uniform	0.1	74	407	4.7	14%

Random Projection

Random Projection to dimensions	Tokens Retrieved	Total Tokens	Deviation in cosine similarity	% reduction in attack success
Baseline	93	407	-	-
768	60	407	0.7	35%
768+Noise	54	407	5.9	42%
576	41	407	0.73	55%

Random Rotation

Method	Tokens Retrieved	Total Tokens	Deviation in cosine-similarity	% reduction in attack success
Baseline	73	407	-	-
Random Rotation	72	407	1.37	1.3%
Random Rotation+ Noise	62	407	6.49	15%
Random Rotation+Random Projection	37	407	1.23	49%

Squaring Embeddings

Method (Embedding dimensions Squared)	Tokens Retrieved (After)	Total Tokens	Deviation in cosine-similarity	% reduction in attack success
Baseline	86	407	-	-
192	61	407	0.32	30%
576	74	407	1.03	14%
768	81	407	2.15	6%

Observations:

- **Random Projection** provided relatively the best defense with minimal impact on similarity.
- **Simple Gaussian noise** improved privacy modestly but was insufficient alone.
- **Combining rotation and projection** achieved nearly **50% attack reduction**.
- **Over-aggressive dimension squaring** harms cosine preservation and privacy.

Extended Experiments:

- Longer attack training still preserved the observed defense trends.
- Performance consistent across different datasets.

9. DP-Forward

DP-Forward is a new privacy technique where noise is added during the forward pass of a language model (like BERT) — not during backpropagation like DP-SGD.

It perturbs the embeddings directly (before fine-tuning or inference) to achieve local differential privacy (LDP).

This protects both training data and inference queries from attacks like:

- Embedding inversion attacks
- Sensitive attribute inference

The noise is added before training, meaning all later computations are automatically private (free post-processing).

Attack scripts were trained for one epoch on STSB Test (100 pairs). Here noise is added in the following layers of architecture. As per the paper, deeper layers provide relatively more protection as compared to the first or end layers.

Layer	Tokens retrieved	Total tokens	% of reduction in attack success
1	65	372	81%
7	47	372	87%
10	87	372	77%

Why We Did Not Use DP-Forward

- Architectural Changes:
DP-Forward modifies BERT's internal forward pass, breaking compatibility with

standard pretrained models and complicating deployment.


- Loss of Pretraining Benefits:
Altering the forward computations can disrupt the carefully learned representations from pretraining, harming downstream performance.
- Higher Computational Cost:
Injecting noise during the forward pass increases both training and inference overhead, making the system heavier and slower.
- Our Focus: Lightweight, Post-Facto Defenses:
We aimed for external techniques that require no architectural changes, maintain pretrained benefits, and minimize computational burden.

10. Conclusion

Given the resource demands, computational overhead, and architectural complexity introduced by DP-Forward, we have decided to move forward with external perturbation mechanisms, which offer a much more feasible path for deployment. Among these, Random Projection consistently demonstrates strong performance. Importantly, we find that maintaining the original 768-dimensional embedding space — rather than reducing dimensionality — is sufficient to achieve up to a 50% reduction in attack success rate. This consistency is further validated through the attached experimental results, which show stable performance across multiple training epochs and different datasets where we can notice as much as 65% reduction in attack success rate.

11. Appendix

1. Results for attack results for increased training epochs, different dataset

 Experiments Results

2. DP-Forward paper:  IP PAPER.pdf