Differentially Private Text Generation for Authorship Obfuscation using ER-AE

Nico Manzonelli and Joeseph Zuccarelli

HARVARD School of Engineering and Applied Sciences

CS208: Applied Privacy and Data Science

Adversarial Scenario

There are many cases in which the author of some text document wishes to remain anonymous. In the past, authors could simply publish under a pseudonym to obscure their identity. However, with recent developments in the field of natural language processing, an adversary could use computational authorship attribution methods can to re-identify anonymous text documents. The adversary faces an authorship verification problem. We propose an authorship verification attack which is similar to a re-identification attack. In a re-identification attack, the adversary attempts to re-identify an individual's private data by using an auxiliary source. Authorship verification attacks follow a similar format. In an authorship verification attack, we assume the adversary holds a piece of text from a known author (the auxiliary source) and attempts to verify if a corresponding piece of text has the same author (private data).

Connection to Differential Privacy

As a defense to author attribution, researchers study authorship obfuscation, which uses computational methods to release text while concealing enough information to protect against authorship attribution. However, many obfuscation techniques rely on empirical evaluation and provide little to no theoretical guarantees. Therefore, the main question becomes how can we release anonymized text documents using differential privacy (DP)?

Previous Work

- Originally tackled by Weggenmann and Kerschbaum, who focused on creating differentially private term-frequency vectors. They define many key DP principles in terms of text (SynTF, 2018).
- Let x be a text document which could be a sentence, paragraph, or entire article in dataset X. x is a feature vector consisting of words from a fixed vocabulary set V.
- To define adjacency, we consider two datasets X and X' that each have at least a single observation. X is adjacent to X' if any edits are made to the document in X.
- Sensitivity is bounded by the use of a rating function, ρ, which maps an input document, x, an output document, z, in $\mathbb{R}$. The sensitivity is $\max_{z \in Z} \max_{x \sim x'} \rho(x,z) - \rho(x',z)$. If we bound ρ to [0,1] then the worst-case sensitivity is 1.
- Given the privacy parameter, ε, and the length of all text documents bound by l, a text generation function is differentially private if for each $\{x_i, \dots, x_l\} = x \in X$ the generation function outputs $z = \{z_i, \dots, z_l\}$ such that each z_i satisfies $(\varepsilon, 0)$ −DP. Therefore, the generated text document is $l(\varepsilon, 0)$ −DP. via simple composition.

State of The Art – ER-AE

The *Embedding Reward Auto Encoder,* ER-AE, model proposes using a seq2seq autoencoder with a two-set exponential mechanism and custom *REINFORCE* reward function to generate differentially private text. The two set exponential mechanism used to sample from the PMF generated by the decoder (generator) is $l(\varepsilon + \ln(s))$ −DP, where s is the the size of the vocabulary set V. The *REINFORCE* reward function rewards the model for learning to assign higher probabilities to semantically similar tokens. The encoder applies two layers of stacked bi-directional GRUs to make the latent space. The decoder passes the latent space to another NN with two layers of staked bi-directional GRUs. For each $z_i \in z$, the decoder calculates a logit weight for every candidate token $v \in V$, applies the softmax function to generate $\Pr[z\_i = v]$, and uses the resulting probability mass function to choose a token set, T_i, with the two-set exponential mechanism. Finally, by randomly sampling a token from T_i for each $z_i$the algorithm generates a private text document (Bo et. al, 2021).

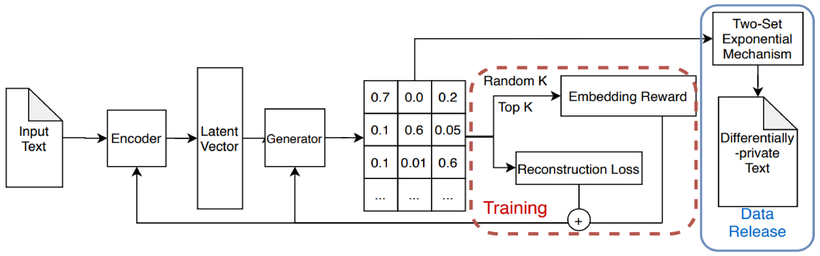

Figure 1: Overall architecture of ER-AE.

Study Objectives

In this project, we implement the ER-AE model to generate differentially private text for authorship obfuscation. We aim to test the robustness of the ER-AE model by evaluating it's performance on a new dataset. More specifically, we systematically feed the model long-text documents with a clear method of privacy accounting and evaluate the model's results in terms of obfuscation performance and semantic preservation while varying the privacy parameter and the dimensionality of the word-embeddings used in training.

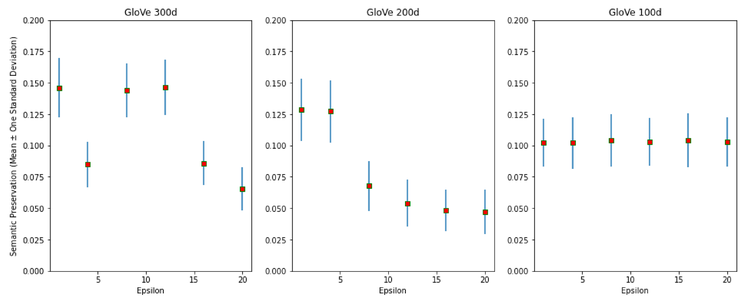

Figure 2: Semantic Preservation Performance

Implementation and Evaluation

- Data: we evaluate the model's performance on long-text documents from the PAN dataset for authorship verification.
- Instead of increasing l to fit the desired length for each long document, we break down each long document into a series of short documents by sentence and recombine them after applying ER-AE. Via the composition theorem of differential privacy, the resulting document is $n * l * (\epsilon + \ln(s))$-DP, where n *is* the number of sentences with length $\leq l$ in the document at hand.
- Evaluate model performance in terms of authorship obfuscation (decrease auc score from baseline authorship verification model), and semantic similarity (cosine distance between sentence semantic embedding generated using Google's Universal Sentence Embedding model).

Results

- The 300-dimensional embeddings typically capture more of the original message (each use the terms 'gaze' and 'ocean'). However, the sentences in Table 1 are not great.
- Figures 2 and 3 highlight that higher dimensional typically perform better at obfuscating the author or preserving semantic meaning. Note that we do not see a clear change in authorship obfuscation performance across epsilons likely due to privacy accounting issues.

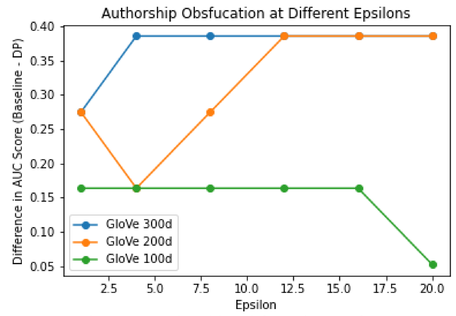

Figure 3: Authorship Obfuscation Performance

Limitations and Future Work

- It is very computationally costly to train and evaluate our model, which limited the interpretability of results.
- To account for privacy accounting issues, we could adaptively adjust l by training many models for each sentence.
- Train the decoder using DP optimization methods (DP-SGD) and compare privacy loss.

Original Sentence: But instead of heeding her warning, he instead turns his gaze back to the ocean.

ϵ	GloVe 100d	GloVe 300d
1	wills respectful of research of ignoring activated stops undertake dimed eyes back to cherubs air sunlight	ravens vibrations dividing dropped her random owls he kay at his gaze back shouldered the ocean
5	strategy instead of bother his l chips he disregard turning sherlocks gaze back to the obvious	junior of many pitch dulled computer steadfast he turns turns his gaze back to the ocean
10	plainly bred of travel his goblin aggravation he impossibly imitate projects gaze back to the cave surface	ravens unacceptable of icy her warning assemble he turns villains her gaze back to the ocean

Table 1: Differentially Private Generated Text Examples