

# ER-AE: Differentially Private Text Generation for Authorship Anonymization

Haohan Bo <sup>1</sup>   Steven H. H. Ding <sup>2</sup>   Benjamin C. M. Fung <sup>1</sup>  
Farkhund Iqbal <sup>3</sup>

<sup>1</sup>McGill University <sup>2</sup>Queen's University <sup>3</sup>Zayed University

May 11, 2021

# Background

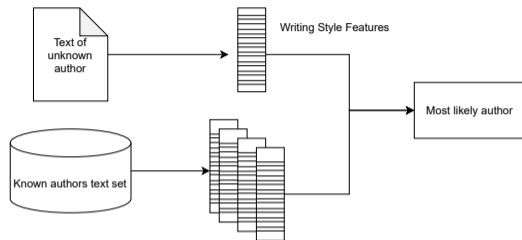


Figure: Authorship Identification.

- Privacy is a vital issue in online data gathering and public data release. However, the studies on privacy protection for textual data are still preliminary.
- Most related works only focus on replacing the sensitive key phrases in the text (Vasudevan and John, 2014) without considering the author's writing style, which is indeed a strong indicator of a person's identity. Stylometric techniques can identify an author of the text from 10,000 candidates based on writing style.

# Background

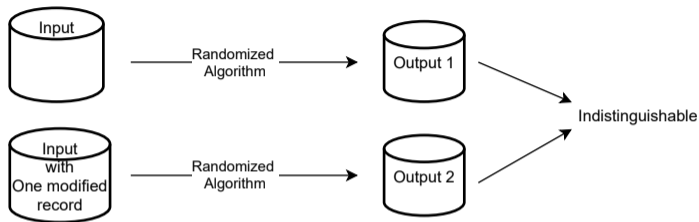


Figure: Differential Privacy.

Differential privacy has received a lot of attention in the machine learning community. It protects the privacy of individual records by achieving the indistinguishability of a single record among other records in the whole dataset. Some work have shown promising results on privacy-preserving text mining. However, their can only output numeric term vectors. Incorporating text generation models with differential privacy mechanism can protect the text privacy by achieving text indistinguishability so that one can hardly recover the original author's identity.

# Challenges

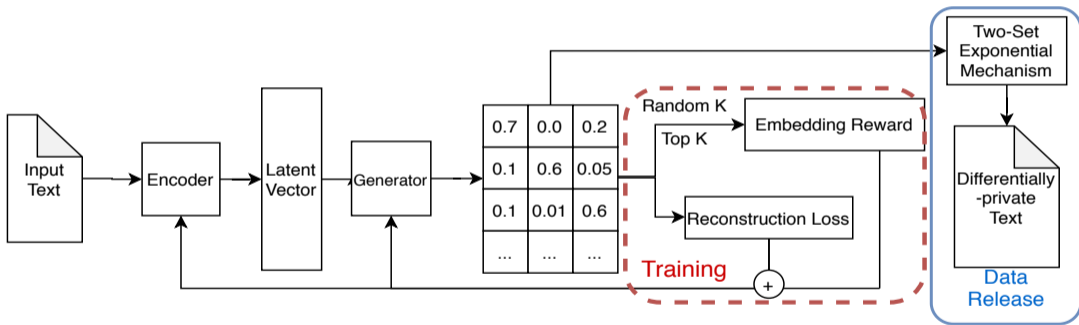
It is challenging to combine a differential privacy mechanism with text generation models. Differential privacy mechanism protects individual records through a randomized algorithm. Because textual data is discrete:

- It's nontrivial to keep semantic information and grammatically correct structure under randomized algorithms. A small movement in the distribution could result in generating a word with totally different meaning.
- Text generation tasks usually have a very large output space (vocabulary), but existing differential privacy mechanisms do not work well on large discrete space.

# Contributions

- The first differentially private authorship anonymization model that can generate human-friendly text in natural language, instead of a numeric vector.
- A novel two-set exponential mechanism to overcome the large output space issue while producing meaningful results.
- A novel combination of a differential privacy mechanism with a sequential text generator, providing a privacy guarantee through a sampling process.
- A new REINFORCE reward function that can augment the semantic information through external knowledge, enabling better preservation of the semantic similarity in the data synthesis process.
- Comprehensive evaluations on two real-life datasets, namely *NeurIPS & ICLR peer reviews* and *Yelp product reviews*, show that ER-AE is effective in obfuscating the writing style, anonymizing the authorship, and preserving the semantics of the original text.

# Embedding Reward Auto-Encoder (ER-AE)



# Generation Procedure of ER-AE

**Input:** Text:  $x$ , Parameters:  $\theta$ , Encoder:  $E_\theta()$ , Generator:  $G_\theta()$ , Privacy budget:  $\epsilon$ .

Produce the latent vector:  $E_\theta(x)$ .

Get probabilities of new tokens:  $Pr[\tilde{x}] \leftarrow G_\theta(E_\theta(x))$ .

**FOR**  $i \leftarrow 1$  to length of  $x$  **DO**

    Build two candidate token sets based on  $Pr[\tilde{x}_i]$ :  $S$ ,  $O$ .

    Apply exponential mechanism to choose token set:  $T$ .

    Randomly sample new  $i$ -th token from  $T$ :  $\tilde{x}_{dp}[i]$ .

**ENDFOR**

**Output:** Differentially Private Text:  $\tilde{x}_{dp}$ .

# Preliminary: Differential Privacy

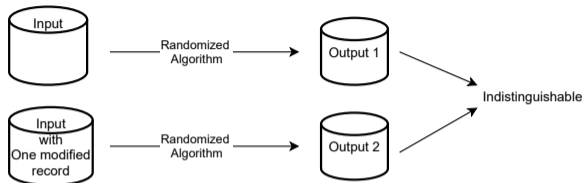


Figure: Differential Privacy.

## Definition

**Differential Privacy.** Two datasets are considered as adjacent if there is only one single element is different. Let privacy budget  $\epsilon > 0$ , a randomized algorithm  $\mathcal{A} : D^n \rightarrow Z$ , and the image of  $\mathcal{A}$ :  $im(\mathcal{A})$ . The algorithm  $\mathcal{A}$  is said to preserve  $\epsilon$ -differential privacy if for any two adjacent datasets  $D_1, D_2 \in D^n$ , and for any possible set of output  $Z \in im(\mathcal{A})$ :

$$Pr[\mathcal{A}(D_1) \in Z] \leq e^\epsilon \cdot Pr[\mathcal{A}(D_2) \in Z]$$

□



# Differentially Private Text Generation

## Definition

**Differentially Private Text Generation.** Let  $\mathbb{D}$  denote a dataset that contains a set of texts where  $x \in \mathbb{D}$  is one of them.  $|x|$ , the length of the text, is bound by  $l$ . Given  $\mathbb{D}$  with a privacy budget  $\epsilon$ , for each  $x$  the model generates another text  $\tilde{x}_{dp}$  that satisfies  $\epsilon/l$ -differential privacy.

□

Following the above definitions, any two datasets that contain only one record are probabilistically indistinguishable w.r.t. a privacy budget  $\epsilon$ .

# Differential Privacy for Discrete Data

The *exponential mechanism* (McSherry and Talwar, 2007) can be applied to both numeric and categorical data (Fernandes et al., 2018). However, according to Weggenmann and Kerschbaum (2018), the exponential mechanism requires a large privacy budget to produce meaningful results while the output space is large, the vocabulary size in our case. It's unlikely to randomly sample a good result directly among 20,000 candidates.

# Two-set Exponential Mechanism workflow

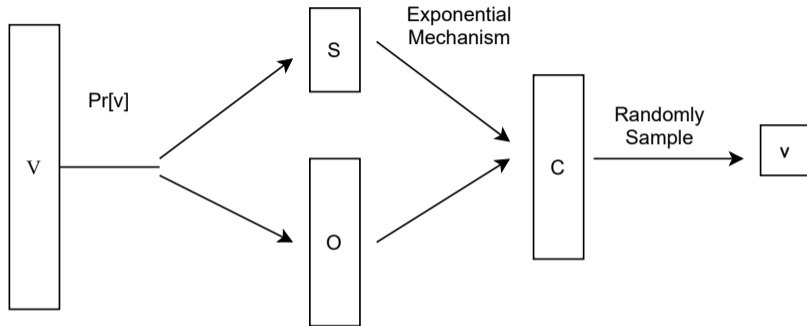
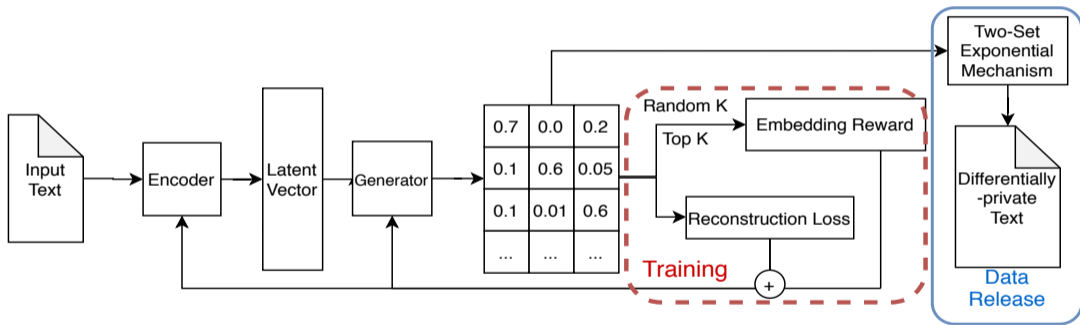


Figure: Two-set Exponential Mechanism workflow.

## Theorem

Two-Set Exponential Mechanism. *Given a privacy budget  $\epsilon > 0$  and the size of output space  $s$ , two-set exponential mechanism is  $(\epsilon + \ln(s))$ -differentially private.*

# Embedding Reward Auto-Encoder (ER-AE)



# Initial Grammar and Semantic Preservation

We follow an unsupervised learning approach since we do not assume any label information. First, we adopt the reconstruction loss function:

$$\mathcal{L}_{recon} = \sum_{x_i \in \mathcal{X}, x \in \mathbb{D}} -\log Pr [\tilde{x}_i = x_i] \quad (1)$$

It maximizes the probability of observing the original token  $x_i$  itself for the random variable  $\tilde{x}_i$ . In the recent controllable text generation models, the reconstruction loss plays an important role to preserve grammar structure and semantics of input data when combined with the other loss.

# REINFORCE Training for Semantic Augmentation

The text dataset to be anonymized and released can be small, and the extra semantic knowledge learned from the other corpus can provide additional reference for our rating function. This reward function is inspired by the Policy Gradient loss function,  $\mathcal{L}_{embed}$  is:

$$- \sum_{x_i \in X, x \in \mathbb{D}} \left( \sum_{v \in \mathbb{E}_k(\tilde{x}_i)} \log(\text{Pr}[\tilde{x}_i = v])\gamma(x_i, v) + \sum_{w \sim \mathbb{V}_k} \log(\text{Pr}[\tilde{x}_i = w])\gamma(x_i, w) \right)$$

At time step  $i$ , this reward function first assigns rewards to the top- $k$  selected tokens, denoted as  $\mathbb{E}_k(\tilde{x}_i)$ , according to probability estimates for random variable  $\tilde{x}_i$ .

We evaluate our model with the following two datasets:

- **Yelp Review Dataset:** All the reviews and tips from the top 100 reviewers ranked by the number of published reviews and tips. It contains 76,241 reviews and 200,940 sentences from 100 authors.
- **Academic Review Dataset:** All the public reviews from NeurIPS (2013-2018) and ICLR (2017) based on the original data and the web crawler provided by (Kang et al., 2018). It has 17,719 reviews, 268,253 sentences, and the authorship of reviews is unknown.

# Experiments Results

Table: Results for each evaluation metric on both datasets.  $\uparrow$  indicates the higher the better.  $\downarrow$  indicates the lower the better.

Model	Yelp (100-author)			Conferences' Dataset	
	USE $\uparrow$	Authorship $\downarrow$	Stylometric $\uparrow$	USE $\uparrow$	Stylometric $\uparrow$
Original text	1	0.5513	0	1	0
Random-R	0.1183	0.0188	62.99	0.1356	65.624
AE-DP	0.6163	0.097	11.443	0.614	9.859
SynTF	0.1955	0.0518	26.3031	0.2161	25.95
ER-AE (ours)	0.7548	0.0979	13.01	0.7424	9.838



# Intermediate Results

Table: The intermediate result of top five words and their probabilities at that the third and the fourth generation steps.

---

	<b>several</b>	<b>unique</b>
AE-DP	<b>several 0.98</b> , those 0.007, some 0.003, various 0.002, another 0.001	<b>unique 0.99</b> , different 0.0001, new 3.1e-05, nice 2.5e-05, other 2.1e-05,
ER-AE	many 0.55, some 0.20, <b>several 0.14</b> different 0.04, numerous 0.03	<b>unique 0.37</b> , great 0.21, amazing 0.15, wonderful 0.1, delicious 0.05



---

# Samples

Table: Sample sentences generated by models.

<b>Input</b>	the play place is pretty fun for the little ones .
Random-R	routing longtime 1887 somalia pretty anatomical shallow the dedicated drawer rosalie
AE-DP	employer play lancaster mute fish fun for wallace little chandler .
SynTF	conditioned unique catherine marquis governing skinny garment hu vivid . insists
ER-AE	the play place is pretty nice with the little ones !
<b>Input</b>	i also ordered a tamarind margarita and it was great .
Random-R	substantial char recommended excavation tamarind coil longitudinal recover verify great ho
AE-DP	intersection also ordered service tamarind drooling scratched denis monkfish motions .
SynTF	carnage spence unsigned also clinging said originated beacon liking strike accomplishments
ER-AE	i also requested a tamarind margarita and it were great .
<b>Input</b>	the manuscript is well written is provides good insight into the problem .
SynTF	ness voice incoming depending entrances somehow priscilla rows romantic oblivious mall
ER-AE	the manuscript is well edited has provides excellent insight into the problem .

# References

-  Frank McSherry and Kunal Talwar. 2007.  
Mechanismdesign via differential privacy.  
*Proceedings ofthe 48th Annual IEEE Symposium on Foundationsof Computer Science (FOCS).*
-  Benjamin Weggenmann and Florian Kerschbaum.2018.  
Syntf: Synthetic and differentially privateterm frequency vectors for privacy-preserving textmining.  
*Proceedings of the 41st ACM Interna-tional Conference on Research & Development inInformation Retrieval (SIGIR).*
-  Natasha Fernandes, Mark Dras, and Annabelle McIver.2018.  
Author obfuscation using generalised differ-ential privacy.  
*arXiv preprint.*

Thank you!