ER-AE: Differentially Private Text Generation for Authorship Anonymization

Haohan Bo [1] Steven H. H. Ding [2] Benjamin C. M. Fung [1] Farkhund Iqbal [3]

[1]McGill University [2]Queen's University [3]Zayed University

Research Problem

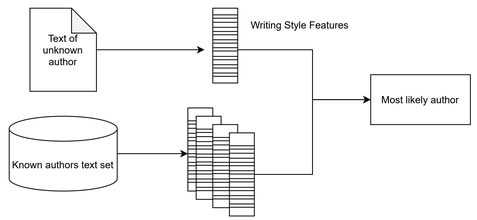

Figure 1. Authorship Identification.

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. Even though some textual data, such as double-blind academic reviews, is released anonymously, the adversaries may recover the author's identity using the personal traits in writing. Stylometric techniques can identify an author of the text from 10,000 candidates based on writing style.

Only a few recent studies focus on authorship anonymization, aiming to hide the personal traits of writing style in the given textual data. However, none of them can produce human-friendly text with a privacy guarantee.

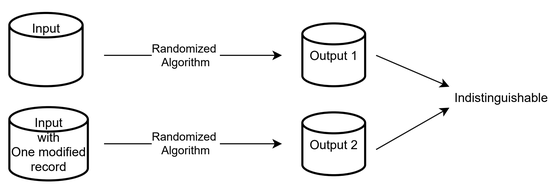

Figure 2. Differential Privacy.

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. 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 mechansim 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 grammarly 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.

Embedding Reward Auto-Encoder (ER-AE)

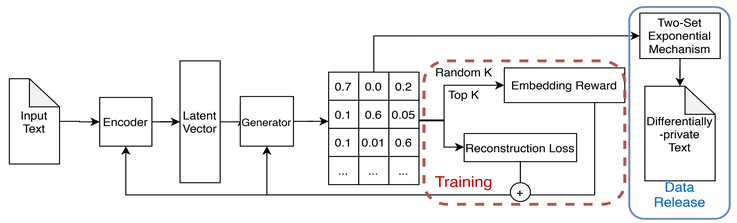

Figure 3. Overall architecture of ER-AE.

Figure 3 depicts the overall architecture of our proposed ER-AE model, which consists of an encoder and a generator. The encoder receives a sequence of tokens as input and generates a latent vector to represent the semantic features. The generator, which is incorporated with the two-set exponential mechanism, can produce differentially private text according to the latent vector. ER-AE is trained by combining a reconstruction loss function and a novel embedding loss function.

Two-Set Exponential Mechanism

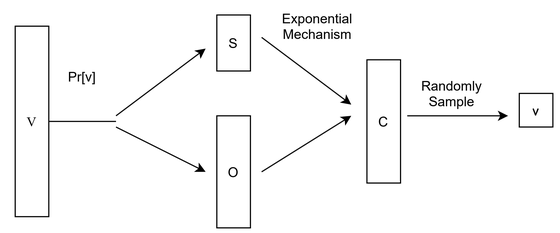

Figure 4. Two-set Exponential Mechanism workflow.

We designed a new differential privacy mechanism to tackle the large outputs space issue. Instead of sampling results from the whole set directly, we build two subsets based on model based probabilities and select one subset through exponential mechanism. The result is sampled randomly from the chosen subset. The two-set exponential mechanism is $(\epsilon + \ln(s))$-differentially private.

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:

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

Datasets

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.

Evaluations

Table 1. Results for each evaluation metric on both datasets. ↑ indicates the higher the better. ↓ indicates the lower the better.

	Yelp (100-author)			Conferences' Dataset	
Model	USE ↑	Authorship ↓	Stylometric↑	USE↑	Stylometric↑
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

Table 2. The intermediate result of top five words and their probabilities at that the third and the forth generation steps.

Input: there are **several unique** hot dog entrees to choose ...		
	several	unique
AE-DP	**several 0.98**, those 0.007, some 0.003, various 0.002, another 0.001	**unique 0.99**, different 0.0001, new 3.1e-05, nice 2.5e-05, other 2.1e-05,
ER-AE	many 0.55, some 0.20, **several 0.14** different 0.04, numerous 0.03	**unique 0.37**, great 0.21, amazing 0.15, wonderful 0.1, delicious 0.05

Table 3. Sample sentences generated by models.

Input	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 !
Input	i also ordered a tamarind margarita and it was great .
Random-R	substantial char recommended excavation tamarind coil longitudinal recover verify great housed
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 .

Acknowledgments

Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (RGPIN-2018-03872), Canada Research Chairs Program (950-232791), and Provost Research Fellowship Award (R20093) and Research Incentive Funds (R18055) from Zayed University, United Arab Emirates. The Titan Xp used for this research was donated by the NVIDIA Corporation.

ben.fung@mcgill.ca