2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) | 978-1-7281-7638-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIIC51459.2021.9415242

Differentially-Private Text Generation via Text Preprocessing to Reduce Utility Loss

Taisho Sasada
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
sasada.taisho.su0@is.naist.jp

Masataka Kawai
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
kawai.masataka.kl6@is.naist.jp

Yuzo Taenaka
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
yuzo@is.naist.jp

Doudou Fall
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
doudou-f@is.naist.jp

Youki Kadobayashi
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
youki-k@is.naist.jp

***Abstract*—To provide user-generated texts to third parties, various anonymization used to process the texts. Since this anonymization assume the knowledge possessed by the adversary, sensitive information may be leaked depending on the adversary's knowledge even after this anonymization. Moreover, setting the strongest assumptions about the adversary's knowledge leads to the degradation of the utility as the data by removing any quasi-identifiers. Therefore, instead of providing original data, a method to generate differentially-private synthetic data has been proposed. Differential privacy is more flexible than anonymization technologies because it does not require the assumption of the adversary's knowledge. However, if a large noise is added to the gradient in text generative model to satisfy differential privacy, the utility of the synthetic text is degraded. Since differential privacy can be satisfied with a small noise in data containing duplicates, it is possible to reduce utility loss as text by creating duplicates before adding noise. In this study, we reduce the amount of noise added by creating duplicates through generalization, thereby minimizing text utility loss. By constructing a differentially-private text generation model, we can provide synthetic text and promote text utilization while protecting privacy information in the text.**

***Index Terms*—Privacy-Preserving Data Mining, Data Privacy, Generative Model, Differential Privacy**

I. INTRODUCTION

User-generated text is one of the most important information for understanding people's attributes. This growth of user-generated text is also observed with the enhanced Consumer Generated Media (CGM) such as blogs, web pages, and message boards. Some organizations provide these texts to third-parties to target new data businesses. Then, it is necessary to remove the identifier to prevent identifying individuals by privacy protection methods.

Privacy protection methods include anonymization through suppression and generalization, respectively [1]. However, sensitive information may be leaked depending on the adversary even after anonymization because most anonymizations must assume the knowledge possessed by the adversary. Therefore, anonymization technologies is not a rigorous method in terms of privacy protection. Differential privacy has been proposed as a method that can address the necessity of assumption [2], [3]. Differential privacy is a mathematical definition for database privacy without the assumption of adversary's knowledge. Differential privacy can handle queries about requiring even whole data while protecting the privacy of individual records in the database. By adding noise to the entire data, it is possible to handle not only queries for mean value and variance but also queries for a part of the data.

With this background, the data publication with differential privacy has been attracting attention, and recently, differentially-private data generation with deep learning has been proposed [4]–[6]. In deep learning, the parameters are updated using stochastic gradient descent (SGD) during repeated training, but if the gradient contains private information during training, it may be a possible to identify individuals from the data generated. To address this threat, we must optimize gradient to protect private information while adding noise to satisfy differential privacy. Hence, it is assumed that we have to conceal private information in gradient by optimizing the text generative model while adding noise. In the case of data providers, they can only publish the synthetic data to a third party while protecting individual's privacy. However, in contrast to numerical data, the text contains many unique words due to the inclusion of named entities such as place names and organization names. This makes it difficult to create duplicates, and the nature of differential privacy requires to add large noise. This makes it difficult to build a model that can generate significant text for data mining.

In this study, we create multiple duplicates by generalizing the named entities before learning, taking advantage of the nature that the differential privacy can be satisfied by adding relatively small noise to the data containing duplicates. In this way, we reduced the amount of noise added to the gradient. By adding relatively small noise to the gradient

during training, we can build a text generative model that satisfies differential privacy and generates synthetic text that cannot identify individuals.

The contributions made by this study are as follows:

- Provision of synthetic text by extending the optimization function for differentially-private generation.
- Suppression of text utility loss via reducing the amount of added noise.
- Enhancement of irreversibility between the original text and generated synthetic text.

The structure of this paper is as follows: First, we give an overview of the related research on differential privacy and its problems in Section II. In Section III, we describe the content of the proposals to address the existing problems. In Section IV, we explain about the contents of evaluation experiments performed to verify the utility and irreversibility of the proposed method and a discussion of the results. In Section V, we summarize this study and refer to future work.

II. RELATED WORK

In this section, we explain the basic preliminaries about differential privacy and the generation of synthetic data for privacy-preserving data mining.

A. (ϵ, δ)-Differential Privacy

Differential privacy depends on the notion of adjacent databases. (ϵ, δ)-differential privacy has been proposed as an extension of differential privacy by Dwork [7]. Here in after, we explain the notation in differential privacy. First, let $D = \{x_1, x_2, \cdots, x_n\}$ be a database with records $x(x \subset \mathcal{X})$ containing personal information. We also define that there is a randomized mechanism $\mathcal{M}$ that returns y $(y \subset \mathcal{Y})$ by adding noise following a specific probability distribution when the database $D(D \subset \mathcal{D})$ is inputted. For two databases D and D', which are adjacent to each other and differ by only one record, then (ϵ, δ)-differential privacy is defined as follows.

Theorem 1: A randomized mechanism $\mathcal{M} : \mathcal{D} \mapsto \mathcal{R}$ gives (ϵ, δ)-differential privacy if for all data sets D and D' on at most one record, and all $\mathcal{S} \subseteq Range(\mathcal{M})$,

$$P[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon) \cdot P[\mathcal{M}(D') \in \mathcal{S}] + \delta \qquad (1)$$

In Equation (1), it is possible to interpret (ϵ, δ)-differential privacy is ϵ-differential privacy except the probability δ. (ϵ, δ)-differential privacy is relaxed, and it defines inequality by adding δ term (see Theorem 1). A randomized mechanism $\mathcal{M}$ satisfying Theorem 1 deals with leakage of private information because no outputs would become significantly more or less likely even if one record is removed from the dataset. This mechanism is based on global sensitivity, which is a measurement of the effect one record can have on a numerical query in the worst case.

Also, from the principle of differential privacy, the privacy loss can be evaluated by the probability to estimate what any given database $\mathcal{D}$ input was when the output o of the randomization mechanism was given.

$$l\left(o; \mathcal{M}, D, D'\right) \triangleq \ln \left| \frac{\Pr[\mathcal{M}(D) = o]}{\Pr\left[\mathcal{M}\left(D'\right) = o\right]} \right| \qquad (2)$$

In Equation (2), $l(o; \mathcal{M}, D, D')$ represents the discriminability between the case with database D and the case with database D'. The larger this probability is, the more identifiable the attacker is, and thus the greater the privacy loss.

B. Generation of Synthetic Data for Privacy Protection

Since any anonymization requires assuming the adversary's knowledge, privacy can be violated depending on the adversary's knowledge. In order to deal with the necessity of assumption, a method to protect privacy is proposed by using generative models to generate synthetic data and providing only the synthetic data generated [5]. Generally, these generative models are trained with deep learning models. Deep learning models update the gradient during training by applying optimizer function but they retain private information that may violate an individual's privacy in this gradient. In recent years, attacks on gradients of deep learning models have been studied, so there is a need to keep these gradients secure and prevent private information leaking.

Therefore, a method to construct a differentially-private model by adding noise to the gradient has been proposed by Shokri and Abadi [4], [8]. In these papers, Gaussian noise is added to optimize the gradient by SGD, which satisfies the (ϵ, δ)-difference privacy instead of pure ϵ-differential privacy. By adding Gaussian noise to the entire record, no single sample will be affected by any one of them.

On the other hand, SGD diverges in the optimization process if the input values contain many unique records. Therefore, to optimize correctly, it is necessary to build a robust model specialized for a specific generative task or to deal with such unique records before training. In this paper, we focus on the text that can be freely described by individuals, and this problem is expected to be faced because the text contains a large number of unique records such as named entities. Since various generative models have been proposed even now and more accurate ones are expected to be proposed in the future, this study proposes not a robust model but a data preprocessing and optimization scheme.

C. Reducing the Amount of Noise in Differential Privacy

Adding a large amount of noise significantly reduces the data utility. To enhance data utility while protecting privacy, Soria proposes a method to satisfy k-anonymity first before adding noise to satisfy differential privacy as an attempt to reduce the amount of noise addition [9]. They created multiple duplicates by generalizing unique values to satisfy k-anonymity before adding noise. In data with many duplicates, it is difficult to distinguish the output when one record is missing from the output when another record is missing. Therefore, they generalized unique values to be duplicated to satisfy differential privacy with relatively small noise.

For named entities in text, it is possible to reduce the amount of noise by anonymizing them to satisfy k-anonymity before

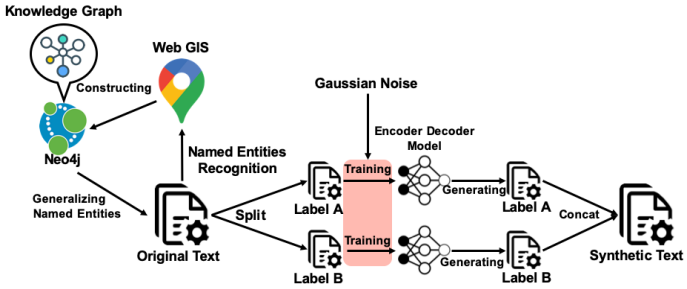

Figure 1. The overall workflow of the proposed method. Original text inputted is generalized by Knowledge Graph [10], and separated based on the label. We construct a generative model based on each label, generate synthetic text, and combine each label's synthetic text.

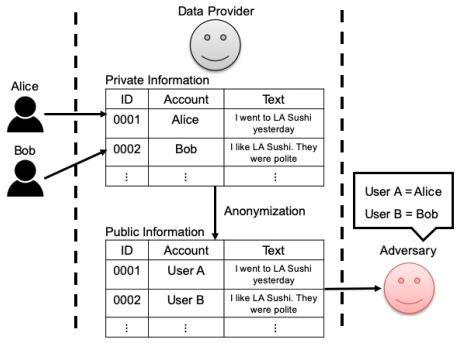

Figure 2. Adversary scenario of this research. Adversary attempts to identify which user wrote the text after anonymization by the data provider.

adding noise in the same way. However, unlike numerical data, the text contains a mixture of generalizable and non-generalizable words. Therefore, it is necessary to first extract generalizable words from the text.

As a method to generalize words in the text, we proposed a method for extracting generalizable named entities, such as place names and organization names, from text and satisfying k-anonymity via a geographic information system, and achieving accurate generalization and notation distortion support considering actual geographic information [10]. In this paper, we use this method to extract and generalize unique expressions in a text to satisfy k-anonymity. We also use this method to extract and generalize the named entities in the text to satisfy k-anonymity and train the generative model while adding noise to the gradient.

III. PROPOSED METHOD

As described in Section II-C, a large amount of noise deteriorates the utility of a dataset. Thus we should generalize words in text because differential privacy can be satisfied with small noise. In this section, we explain how to generalize words in the text and add Gaussian noise to the gradients. Figure 1 shows the overall flow of the proposed method. We first generalize named entities to create duplication in text and then train the generative model while adding Gaussian noise to the gradients. We build a generative model that satisfies differential privacy and attempts to generate the privacy-protected text.

The proposal of this study consists of three procedures: (1) Generalization of named entities in the original text to reduce the amount of noise, (2) Construction of a differentially-private generative model, and (3) Generation of synthetic text based on original text distribution.

Since focusing on the accuracy of polarity classification in experimental evaluation, we assume a binary classification in this synthetic text generation. Thus, we construct a label-by-label optimized generative model.

A. Adversary Scenario

In this section, we describe the adversary scenario. As an adversary scenario, we assume a semi-honest model where an adversary tries to steal the private information of others while following a defined protocol. Figure 2 shows an example of how an adversary uniquely identifies a user and guesses which user wrote the text since the data provider published the anonymized text. In Figure 2, An adversary attempt to identifying the user's attribute information with the written text. For example, if a written text contains travel plans, any personal information, such as the user's address and organization, can be inferred from a previously written text.

B. Generalizing Named Entities in the Original Text

As we described in Section II-C, it is possible to reduce the amount of noise added to satisfy differential privacy and prevent the loss of usefulness. However, in the case of text, it is first necessary to select the words that have superordinate or subordinate concepts and generalize these. According to previous studies, the names of people, organizations, and places are often considered as anonymizing targets because they contribute to identifying individuals [11]. Named entities are general terms for a systematic collection of proper nouns such as the names of people, places, and organizations. In this study, we generalize these three types of named entities.

As we described in Section II-C, it is possible to generalize the names of places and organizations by using knowledge graph. However, people's names do not feature in our proposal [10]. If we apply that generalization method in this study, people's names are left unanonymized. Therefore, we explain how this study generalizes people's names.

We generalize the names extracted by a named entity recognizer until k-anonymity is satisfied. If k-anonymity cannot be satisfied even in the initials-only state, it can be replaced by an asterisk. Also, we used Bidirectional Encoder Representations from Transformers (BERT) as named entity recognizer because BERT has the highest accuracy in any unstructured text [12].

C. Training of a Differentially-Private Generative Model

As we described in Section II-B, providing the original text carries the risk to identify individuals. Thus, this study also proposes a method of providing text generated by a differentially-private generative model. In this section, we describe how to optimize the gradient of the generative model with Gaussian noise to satisfy differential privacy. Our model

Algorithm 1 Optimizing with Differentially-Private ADAM

Input: $\mathcal{X}, \mathcal{J}(\cdot), \eta, \sigma, \mathcal{L}, \mathcal{C}$
Output: θ_T

1: Initialize θ_0, m_0
2: **for** $t \in [T]$ **do**
3: select random sample L_t where sampling probability $\frac{L}{N}$
4: **for** each $i \in L_t$ **do**
5: $g_t(x_i) \leftarrow \nabla_{\theta_t}\ \mathcal{J}(\theta_t, x_i)$
6: **end for**
7: $\overline{g}_t(x_i) \leftarrow \max\ (1, \frac{\|g_t(x_i)\|_2}{C})$
8: $\tilde{g}_t \leftarrow \frac{1}{L}\sum_i(\overline{g}_t(x_i)) + \mathcal{N}(0, \sigma^2 C^2 I)$
9: $m_{t+1} = \beta_1 m_t + (1-\beta_1)\,\tilde{g}_t$
10: $v_{t+1} = \beta_2 v_t + (1-\beta_2)\,\tilde{g}_t^2$
11: $b_{t+1} = \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}}$
12: $\theta_{t+1} = \theta_t - \alpha_t \frac{m_{t+1}}{\sqrt{v_{t+1}}+\phi} b_{t+1}$
13: **end for**

is an Encoder-Decoder model, and we adopt Long Short-Term Memory (LSTM) [13] as an internal structure because LSTM is capable of generating variable-length sentences.

The gradient in the generative model must be protected because this also contains private information by training. Trained models often are released to third parties, but membership inference attacks [14], [15], whose training data can be inferred from output values and gradients, have been considered as threats. Therefore, it is dangerous to provide output values and gradients without privacy protection. In this study, we add Gaussian noise to the gradient during training to satisfy differential privacy.

Algorithm 1 describes the optimization process in our proposal. First, we initialize the parameters θ in Line 1 of Algorithm 1. In Line 3-6, we select sample L_t under sampling probability L/N and compute the gradient at each sample. And we clip the Euclidean norm of the gradient in Line 7-9, we caliculate the average with Gaussian noise to construct a differentially-private generative model. Also, we adopt recommended Adam's parameters: α = 0.001, β_1 = 0.9, β_2 = 0.999, $\phi = 10^{-8}$ [16].

D. Generation of Differentially-Private Text

To generate the text by our model, we need to give parameters for what words to generate and the length of the text to generate. If we input a word at this point, it may end up with a pseudo-text set that is completely different from the original text set. Therefore, it is necessary to generate them based on the probability of occurrence of the words in the original text set. In Algorithm 2, we compute word occurrence probability $P(T_w|p)$ in Line 1. In Line 2-5, we select a random token T_w and input T_w to the differentially-private generative model $\mathcal{G}(\cdot)$ because this random sampling makes the synthetic text follow the distribution of original text. Also, $P(T_w|p)$ can be caliculated in Equation (3).

Algorithm 2 Generation of Differentially-Private Text

Input: $\mathcal{X} : \{x_1, x_2, \cdots x_n\}, \mathcal{G}_\theta(\cdot), l_i$
Output: $\mathcal{X}^{dp}$

1: compute $P(T_w|p)$
2: **for** $w \in [d]$ **do**
3: select token T_w where sampling probability $P(T_w|p)$
4: generate differentially-private text x_i^{dp} by $\mathcal{G}_\theta(T_w, l_i)$
5: add x_i^{dp} differentially-private text set $\mathcal{X}^{dp}$
6: **end for**
7: **return** $\mathcal{X}^{dp}$

$$P(T_w|p) = \prod_{w=1}^{n} p^{T_w}(1-p)^{1-T_w} \tag{3}$$

In Equation (3), $T_1, T_2, \cdots T_n$ follows a binomial distribution B(1,p) and can be transformed as $P(T_w|p) = p^{T_w}(1-p)^{1-T_w}$, which leads to express the total power of $p^{T_w}(1-p)^{1-T_w}$. Based on this occurrence probability, we implement probability sampling tokens among original text.

IV. Experiment

In this section, we describe the datasets to evaluate the utility and irreversibility between original text and generated text, evaluation metrics, experimental results, and discussion.

A. Experimental Setup

The outline of the experiment is as follows: First, the original text is divided into train text and test text. Second, we generalize the named entities in the train text and preprocess the text to include duplicates so that the differential privacy can be satisfied with a relatively small noise. Third, we then vectorize the text to construct a differentially-private generative model. In the end, we then generate privacy-protected text from the differentially-private generative model and use the generated text to evaluate the utility loss and irreversibility.

To evaluate utility loss of generated text by our proposal, and irreversibility to inference training text, we use IMDB Movie Review Dataset [17], which is commonly used for polarity classification. This dataset contains 100,000 movie reviews and is categorized as either negative or positive. For the hyperparameters of the differentially-private generative model, we set it as follow; the epoch is 500 in training, the dropout rate in LSTM is 0.2, and we used ReLU as the activation function in the hidden layer and Softmax function as the activation function in the output layer, respectively. Also, we adopted cross-entropy for the model compilation loss function and Adam for the optimization function to avoid being trapped in local optimization where the model generates only certain words.

B. Evaluation of Utility Loss as Text for Privacy Protection

The evaluation of the utility is based on the classification of positive and negative labels. The smaller the difference of results between original text and generated text, the closer the

Table I
UTILITY LOSS : RESULTS OF POLARITY CLASSIFICATION

	Original Text		DP E-D		**DPvG E-D(Ours)**	
	POS	NEG	POS	NEG	POS	NEG
MLR	0.582	0.681	0.426	0.622	0.510	0.648
SVM-RBF	0.642	0.723	0.577	0.653	0.609	0.680
MNB	0.640	0.755	0.595	0.661	0.614	0.691
DT-C4.5	0.609	0.739	0.536	0.639	0.570	0.673

learning results are to the original text. In this experiment, we built four different models for classification: Multinomial Naïve Bayes (MNB), Support Vector Machine with RBF Kernel (SVM-RBF), Multinomial Logistic Regression (MLR), and Decision Tree with C4.5 (DT-C4.5). Table I shows F_1-score as evaluation metrics by using these classifier models. In Table I, "Original Text" means the classification results by training of original text without Encoder-Decoder model, DP E-D means the classification results by training differentially-private text via the Encoder-Decoder model, and DPvG means the classification results by training of differentially-private text via generalization and Encoder-Decoder model. This result was obtained when the privacy budget was 1.0, which gave the highest classification accuracy for polarity classifications.

As experimental results, Table I shows that DPvG E-D is more useful than simple DP E-D because the F1-Score is more approximate to the one with normal LSTM in the polarity classification. This is because the generalization of named entities contributes to the generation of low-frequency words and improved variance in text generation while avoiding being trapped in local optimization where generates the same words. As an example of the improved variance by the proposed method, the word "New York," which was not included in the words generated by DP E-D, was included. This may be due to the fact that the frequency of the word increased as a result of generalization from "Brooklyn." On the other hand, for polarity classification, the generalized names of persons, places, and organizations had little effect on the classification accuracy before training, and thus did not lead to a utility loss. However, they may affect the classification task where they make a large contribution. In addition, the sampling of words to be input to the generative model is based on the probability distribution of the original text. Therefore, if the original text contains many same words, the generative model may generate the same text. Generating same text may bias the model, which cause to train on only certain words. Since this bias leads to the deterioration of Utility Loss, it is necessary to diversify the variance of words in the generated text.

C. Evaluation of irreversibility Against Original Text

In the irreversibility evaluation, we evaluate the cosine similarity between the original and the generated text and privacy loss. cosine simirality is a method to measure similarity among the text. If the generated text is similar to the original text, it implies that an adversary can estimate the original text by using the generated text. The higher cosine similarity is, the more predictable the original text can be inferred from the generated text.

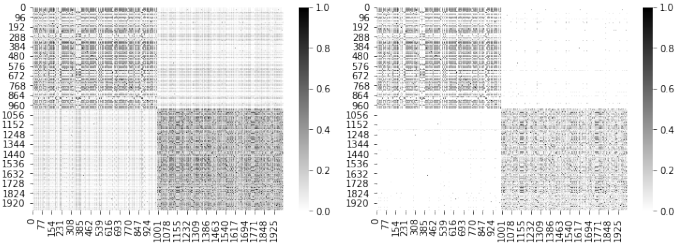

Figure 3. Left: Heatmap of cosine simirality by Differentially-private generative model without generalization before training. Right: Heatmap of cosine simirality by our Differentially-private generative model with generalization before trainig.

$$\mathcal{M}_{xx^{dp}} = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_M \end{array} \begin{array}{c} \begin{array}{ccc} x_1^{dp} & \cdots & x_M^{dp} \end{array} \\ \begin{pmatrix} \cos\theta_{1,1} & \cdots & \cos\theta_{1,M} \\ \cos\theta_{2,1} & \cdots & \cos\theta_{2,M} \\ \vdots & \ddots & \vdots \\ \cos\theta_{M,1} & \cdots & \cos\theta_{M,M} \end{pmatrix} \end{array} \tag{4}$$

$$, \cos\theta_{i,j} = \frac{\mathcal{M}_{xx^{dp}}[i,*] \cdot \mathcal{M}_{xx^{dp}}[*,j]}{\left\|\mathcal{M}_{xx^{dp}}[i,*]\right\| \left\|\mathcal{M}_{xx^{dp}}[*,j]\right\|}$$

The cosine similarity matrix $\mathcal{M}_{xx^{dp}}$ is a matrix whose elements are the cosine similarities between the original document d and the generated document x^{dp}. In this case, documents x and x^{dp} are Residual-IDF [18] weighted to compute the cosine similarity in order to reduce the weight of each word's probability of occurrence, respectively. The TF-IDF, which calculates the importance of a word from two pieces of information, Term Frequency (TF), which is the frequency of occurrence of the word, and Inverse Document Frequency (IDF), which is the inverse document frequency of the word, is the most common way to vectorize such documents. However, TF-IDF is affected by the difference in the number of words per document. Therefore, we use Residual-IDF for vectorization in this study. We lower the weight of these distinctive words because distinctive words such as named entity have a significant impact on privacy leakage. Residual-IDF is a weighting method based on Poisson distribution, which reduces only the weights of words that appear commonly in the documents, without reducing the weights of characteristic words.

Figure 3 shows the heatmap of cosine similarity. There is a difference in the cosine similarity matrix with and without adjustment for the amount of Gaussian noise, indicating that the text with Gaussian noise adjustment is less similar to the original. This is because the generalization of the named entities to reduce the amount of Gaussian noise, which does not generate named entities in the original text, makes the cosine similarity smaller than in the usual differentially-private model.

Moreover, we need to measure the amount of privacy loss caused by the decrease in the amount of added noise, we calculated the privacy loss as the maximum probability of

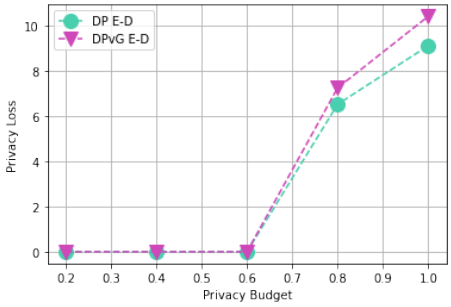

Figure 4. Line graph illustrating the relationship between privacy budget and privacy loss, and DP E-D is the differentially-private encoder-decoder model without the proposed method, DPvG E-D is the encoder-decoder model with our proposed method.

the output word when each word is the input by following Equation (2). Figure 4 shows the relationship between privacy budget and privacy loss. The experimental results showed that DP E-D suppressed the privacy loss more than DPvG E-D. The reason for this is that the DP E-D has more noise added and the output is more biased. The amount of noise added to the DPvG E-D was less than that of the DP E-D, which is thought to have resulted in a larger privacy loss. The privacy loss of DPvG E-D was larger than that of DP E-D because the amount of added noise was smaller. In other words, this index evaluates randomness, and DPvG E-D, the proposed method, was able to generate more diverse words.

D. Discussion of Results

These results suggest that the data could be made irreversibility without diminishing their utility for the task of polarity classification. Also, our method can be applied for other natural language tasks such as machine translation, automatic summarization, and online search engine because our proposal did not focus on polarity classification.

On the other hand, a limitation of this study is that it is difficult to apply to tasks that are heavily influenced by the names of places, organizations, and people. By generalizing these named entities to reduce the added Gaussian noise, it is assumed that the utility as text cannot be guaranteed if these named entities directly affect a task.

V. CONCLUSION

In this paper, we propose a framework for providing useful data while protecting privacy by constructing a differentiall-private generation model that generates text, rather than providing text directly and providing only the generated text. The evaluation experiments show that it is possible to generate more useful text than the usual differentially-private model and protect privacy.

As future work, there are three points that need to be addressed: The first is to examine the generation using transformer models such as BERT. The second is to reduce the Gaussian noise by generalizing other named entities as well, and the last one is to develop more conditional generative models to handle multi labels in the future.

ACKNOWLEDGEMENTS

We thank Shuyo Nakatani from Cybozu Labs, Inc. for advice on writing this paper. Also, this work was supported in part by the ICS-CoE Core Human Resources Development Program, and in part by JSPS KAKENHI Grant Number JP18H03234.

REFERENCES

[1] L. Sweeney, "Achieving k-anonymity Privacy Protection using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, 2002.

[2] C. Dwork, "Differential Privacy: A Survey of Results," in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, Springer, 2008.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.

[4] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

[5] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data Synthesis based on Generative Adversarial Networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, 2018.

[6] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially Private Synthetic Data and Label Generation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 98–104, IEEE, 2019.

[7] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data, Ourselves: Privacy via Distributed Noise Generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer, 2006.

[8] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

[9] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing Data Utility in Differential Privacy via Microaggregation-Based *k*-Anonymity," *The VLDB Journal*, vol. 23, no. 5, pp. 771–794, 2014.

[10] S. Taisho, T. Yuzo, and K. Youki, "Anonymizing Location Information in Unstructured Text Using Knowledge Graph," *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications and Services*, vol. 2020, pp. 163–167, 2020.

[11] N. Mamede, J. Baptista, and F. Dias, "Automated Anonymization of Text Documents," in *2016 IEEE Congress on Evolutionary Computation*, pp. 1287–1294, IEEE, 2016.

[12] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy*, pp. 3–18, IEEE, 2017.

[15] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership Inference Attacks against Generative Models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, 2019.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[17] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

[18] K. Church and W. Gale, "Inverse Document Frequency (IDF): A Measure of Deviations from Poisson," in *Natural language processing using very large corpora*, pp. 283–295, Springer, 1999.