Privacy Preserving Text Representation Learning

Ghazaleh Beigi, Kai Shu, Ruocheng Guo
Computer Science and Engineering, Arizona State University
{gbeigi,kaishu,rguo12}@asu.edu

Suhang Wang
College of Information Sciences and Technology, Penn State University
swz494@psu.edu

Huan Liu
Computer Science and Engineering, Arizona State University
huan.liu@asu.edu

ABSTRACT

Online users generate tremendous amounts of textual information by participating in different online activities. This data provides opportunities for researchers and business partners to understand individuals. However, this user-generated textual data not only can reveal the identity of the user but also may contain individual's private attribute information. Publishing the textual data thus compromises the privacy of users. It is challenging to design effective anonymization techniques for textual information which minimize the chances of re-identification and does not contain private information while retaining the textual semantic meaning. In this paper, we study this problem and propose a novel double privacy preserving text representation learning framework, DPText. We show the effectiveness of DPText in preserving privacy and utility.

ACM Reference Format:
Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. Privacy Preserving Text Representation Learning. In *30th ACM Conference on Hypertext and Social Media (HT '19), September 17–20, 2019, Hof, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3342220.3344925

1 INTRODUCTION

Textual information is one of the most significant portions of data that users generate by participating in different online activities such as leaving online reviews and posting tweets. On one hand, textual data consists of abundant information about users' behavior, preferences and needs which is critical for understanding them. For example, textual data has been used by service providers to track users' responses to products and provide them with personalized services. On the other hand, publishing intact user-generated textual data makes users vulnerable against privacy issues. The reason is that the textual data itself contains sufficient information that causes the re-identification of users in the textual database [1] and the leakage of their private attribute information [3].

These privacy concerns mandate data publishers to protect users' privacy by anonymizing the data before sharing it. However, traditional privacy preserving techniques such as k-anonymity and differential privacy are inefficient for user-generated textual data because this data is highly unstructured, noisy and unlike traditional documental content, consists of large numbers of short and informal posts. Moreover, these solutions may impose a significant utility loss for protecting textual data as they may not explicitly include utility into their design objectives. It is thus challenging to design effective anonymization techniques for user-generated textual data which preserve both privacy and utility.

To address the aforementioned challenges, we propose a double privacy preserving text representation learning framework, called DPText. The proposed framework seeks to learn a privacy preserved text representation so that 1) any potential adversary cannot infer whether or not a target text representation is in the dataset, 2) the adversary cannot deduce users' private attribute from the learned representation, and 3) the semantic meaning of the original textual information is preserved. The learned privacy preserved textual information will be then shared with data consumers. Our double privacy preserving framework protects individuals' privacy against re-identification and leakage of private information. Our empirical results show the efficiency of DPText in terms of preserving both privacy and utility of textual data.

2 THE PROPOSED FRAMEWORK-DPTEXT

Our proposed framework, DPText, consists of an auto-encoder for text representation, a differential-privacy-based noise adder, and semantic and private attribute discriminators.

We extract the content representation for a given document using an auto-encoder [4] by minimizing the reconstruction error. Publishing text representation without proper anonymization will let the adversary learn the original text [5], infer if a targeted user's latent textual representation is in the database or which record is associated with it. We deploy the differential privacy technique to add random noise, i.e., Laplacian noise, to the original representation w.r.t. a given privacy budget, ϵ. Besides guaranteeing differential privacy, adding noise minimizes the chance of the text re-identification and original text recovery.

Adding noise to the latent representation not only may destroy the semantic meaning of the text but also does not necessarily prevent leakage of private attribute information from the text data. Semantic meaning of the text data is task-dependant. For example, in the case of sentiment analysis, sentiment is one of semantic meaning in the given text and sentiment prediction is a classification task. Private-attribute information is also another important aspect of user privacy and includes information that the user do not want to disclose such as age, gender, and location. We therefore need to add an optimal amount of noise to the text latent representation. We approach this challenge by automatically *learning* the amount of the added noise with the privacy budget ϵ. We utilize two semantic meaning and private attribute discriminators to infer the amount of the added noise. The semantic meaning discriminator D_S ensures that the added noise does not destroy the semantic meaning w.r.t.

HT '19, September 17–20, 2019, Hof, Germany

ACM ISBN 978-1-4503-6885-8/19/09.
https://doi.org/10.1145/3342220.3344925

a given task. The private attribute discriminator D_P also guides the amount of the added noise by ensuring that the manipulated representation does not include users' private information.

Inspired by the idea of adversarial learning, we achieve our goals by modeling the objective function as a minmax game among the two introduced discriminators, D_S and D_P. Assume that there are T private attributes. Let $\theta_{D_P^t}$ and θ_{D_S} demonstrate the parameters of private-attribute discriminator model D_P and semantic meaning discriminator model D_S, respectively. The correct labels for the t-th sensitive attribute and semantic classification task in n-th document are also represented by $p_{n,t}$ and y_n, respectively. With N documents, we can write the objective function as follows:

$$\min_{\theta_{D_S},\epsilon} \max_{\{\theta_{D_P^t}\}_{t=1}^T} \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{D_S}(\hat{y}_n, y_n) - \alpha \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{D_P^t}(\hat{p}_{n,t}, p_{n,t}) \quad (1)$$
$$+ \lambda\Omega(\theta) \qquad s.t. \qquad \epsilon \leq c_1$$

where c_1 is a predefined privacy budget constraint, $\mathcal{L}_{D_P^t}$ and $\mathcal{L}_{D_S}$ denote cross entropy loss function, $\hat{p}_t$ is the predicted t-th private attribute, $\hat{y}$ is the predicted semantic, $\theta = \{\theta_{D_S}, \epsilon, \{\theta_{D_P^t}\}_{t=1}^T\}$ is the set of all parameters to be learned and $\Omega(\theta)$ is the parameters regularizer. Note that the resultant text representation satisfies $\tilde{\epsilon}$-differential privacy, where $\tilde{\epsilon} \leq c_1$ is the optimal learned budget[1].

3 EXPERIMENTS

To evaluate the utility of data, we report results w.r.t. a well known text-related task, sentiment analysis. To examine privacy of users, we consider different private information, i.e., age, location, and gender, and report results for private attribute prediction task.

Data. We use a dataset from TrustPilot website [6] where users can write reviews and leave a one to five star rating. Each review is associated with three attributes, gender (male/female), age (over-45/under-35/in-between), and location (Denmark, France, United Kingdom, and United States). We follow the same approach as in [7] to preprocess data. We discard all non-English reviews and sample 10k reviews for each location to balance locations. Each review's score is considered as the target sentiment class.

Model and Experimental Settings. For the document auto-encoder, we use single-layer RNN with GRU cell of input/hidden dimension with d=64. For semantic and private attribute discriminators, we use feed-forward networks with single hidden layer with the dimension of hidden state set as 200, and a sigmoid output layer. The parameters $\alpha = 1$ and $\lambda = 0.01$ are determined through cross-validation. The privacy upper-bound constraint is also set as $c_1 = 0.1$ to ensure the ϵ-differential privacy, $\epsilon = 0.1$. We perform 10-fold cross validation. We report accuracy and $F1$ scores for sentiment and private-attribute predictions, respectively. We apply the trained semantic and private attribute discriminators to test data for utlity and privacy evaluation, respectively. DPText is compared with:

- **Original**: This is original text representation from auto-encoder.
- **DifPriv**: This baseline just adds Laplacian noise ($\epsilon = 0.1$) to the original representation.
- **ADV-ALL** [7]: This method utilizes a generator and a discriminator to generate a text representation that has high utility and high privacy for private-attributes.

[1]Details of the proposed framework and proof can be found in [2].

Model	Sentiment (Acc)	Private Attribute (F1) Age	Loc	Gen
Original	0.7493	0.3449	0.1539	0.5301
DifPriv	0.7397	0.3177	0.1411	0.5118
ADV-ALL	0.7165	0.3076	0.1080	0.4716
DPText	0.7318	0.1994	0.0581	0.3911

Table 1: Higher accuracy shows higher utility, while lower $F1$ demonstrates higher privacy.

Experimental Results. The results of sentiment prediction for DPText is comparable to the Original and it outperforms ADV-ALL. This means that DPText preserves the semantic meaning of the textual w.r.t the given task (i.e., high utility). However, DifPriv performs better than DPText and the reason is that DPText applies noise at least as strong as DifPriv. Therefore, adding more noise results in bigger utility loss. The results of private attribute prediction for DPText is significantly better than Original, DifPriv and ADV-ALL (lower $F1$). This indicates the importance of private attribute discriminator and shows that solely adding noise to satisfy ϵ-differential privacy does not protect textual information against leakage of private attributes. These results indicate that DPText can successfully obscure private information.

To recap, DPText has achieved the highest accuracy and thus reached the highest utility in comparison to other methods. It also has comparable utility results to Original. Moreover, DPText has the best results in terms of privacy compared to the other approaches. These findings confirm the effectiveness of our proposed model in terms of both privacy and utility.

4 CONCLUSION

In this paper, we propose a double privacy preserving text representation learning framework, DPText, which learns a text representation that (1) is differentially private, (2) obscures users' private information, and (3) retains high utility for a given task. Our results show the effectiveness of DPText in minimizing chances of privacy leakage while preserving text semantic meaning. One future direction is to generate privacy preserving text (e.g., sentences, paragraphs) which is critical for having interpretable results.

ACKNOWLEDGMENTS

This material is based upon the work supported, in part, by NSF #1614576, ARO W911NF-15-1-0328 and ONR N00014-17-1-2605.

REFERENCES

[1] Ghazaleh Beigi and Huan Liu. 2018. Privacy in social media: Identification, mitigation and applications. *arXiv preprint arXiv:1808.02191* (2018).
[2] Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. I Am Not What I Write: Privacy Preserving Text Representation Learning. *arXiv preprint arXiv:1907.03189* (2019).
[3] Valentina Beretta, Daniele Maccagnola, Timothy Cribbin, and Enza Messina. 2015. An interactive method for inferring demographic attributes in Twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM.
[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[5] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep models under the GAN: information leakage from collaborative deep learning. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*.
[6] Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of WWW*.
[7] Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards Robust and Privacy-preserving Text Representations. (2018).