Quest Journals
Journal of Software Engineering and Simulation
Volume 10 ~ Issue 2 (2024) pp: 23-28
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Personal Information Protection Method from Text Data based on Word Substitution

Shuangjiao Li[1]
[1]*School of Control and Computer Engineering, North China Electric Power University, Baoding, 071051, China*
Corresponding Author: Shuangjiao Li

ABSTRACT: *With the development of information technology, the Internet has become a part of our lives, and people often post personal remarks on blogs, microblogs, social networks, and other platforms. However, these texts contain various personal privacy information such as the age, race, gender, marital status, etc. of the authors, and there is a risk of personal privacy leakage. This paper proposes a textual personal privacy information protection method based on word substitution, which gives a textual substitution scheme with a low risk of privacy leakage through synonym substitution for personal speech or text posted by users, in compliance with the grammar and basically without changing the original semantics. The experimental results show that the method proposed in this paper significantly reduces the accuracy of the prediction model while keeping the text semantics unchanged.*

KEYWORDS: *Privacy Protection, Natural Language Processing, Adversarial Attacks, Deep Learning*

R*eceived 23 Feb., 2024;* ***R****evised 28 Feb., 2024;* ***A****ccepted 07 Mar., 2024* © *The author(s) 2024.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

With the development of information technology, the network has become part of people's living space, forming a unique society, and various social applications have seen explosive growth. WeChat, blogs, microblogs, and other major social platforms, hundreds of millions of users on the platform to publish personal remarks, or to communicate and interact with others, and billions of data are generated every day. In such a big data era, how to mine the new knowledge generated by these data has become a research favorite. Text categorization is an important topic in data mining with a wide range of applications, including spam detection, content auditing, user intent categorization, sentiment analysis, question answering, news categorization, and so on. Along with the development of machine learning models, text categorization algorithms have been constantly revolutionized, and the emergence of deep learning has elevated it to a new level, deep learning-based models have surpassed the traditional machine learning-based methods in various text categorization tasks.

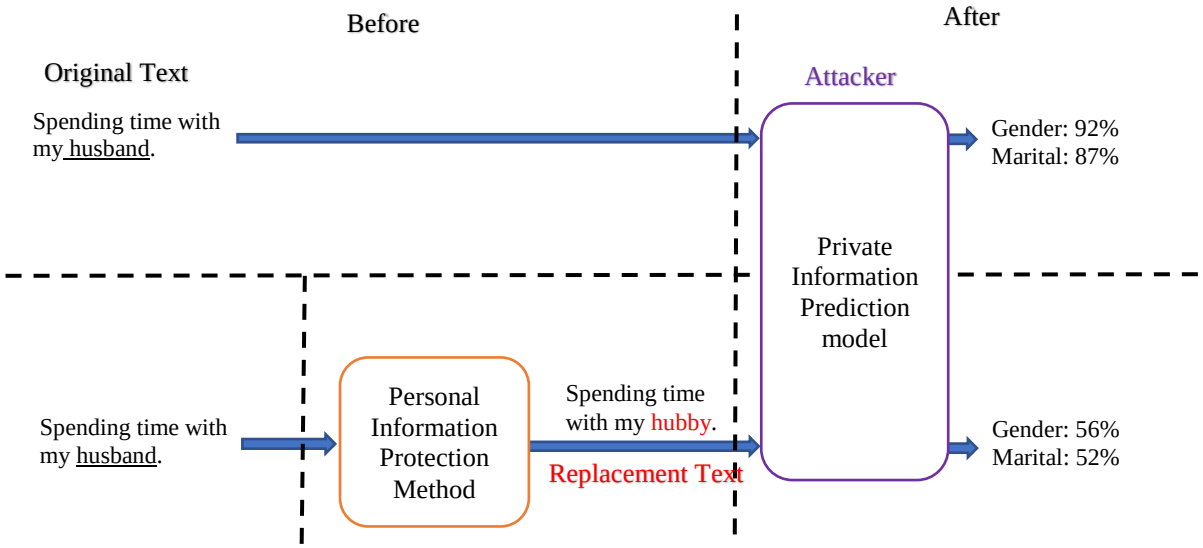
Nowadays, some public personal statements and texts are easily accessible using data collection techniques, and these data contain more private personal information such as the author's personal preferences, character habits, and recent living conditions. A malicious attacker can use the existing public datasets and data labeled with personal information labels such as age, race, gender, marital status, etc., to train on a deep learning model to obtain a text privacy classifier. With the emergence of large language models represented by chatGPT, the models are more capable of prediction and data mining. This also means that large language models trained on big data raise more serious hidden risks of personal privacy leakage[1][2]. It is a serious challenge in the information age to prevent privacy leakage and protect users' privacy while making social platforms fulfill their social functions and enhance people's social connections.

To solve the above problems, previous privacy-preserving research has either trained against the text[3][4] or rewritten the text directly[5][6] in an attempt to eliminate the privacy information in the textual data. Both of these methods require the high-performing adversarial model or recapitulation model, which can be quite costly in terms of defense.

In this paper, we propose a method for protecting personal privacy information in text based on word substitution. The main idea is to perform synonym replacement on the original text to reduce the risk of text

privacy information leakage from the root to achieve the purpose of privacy protection under the premise of complying with the grammar and not changing the original semantics.

Diagram 1 shows the attack and defense process of privacy information, the upper half of the figure shows the process of attackers predicting personal privacy information through deep learning models, the accuracy rate can even be as high as 90% or more. The lower half shows the protection method proposed in this paper. Before a user posts, some words in the original text are synonymously replaced to generate a new



replacement text. When the attacker utilizes the replacement text for prediction, the accuracy of the user's private information prediction will show a significant decrease and the risk of privacy leakage will be reduced.

Diagram 1

The experimental results show that our method can effectively protect the private information hidden in the text. It also ensures that there is no significant difference in the semantics of the text before and after modification. The algorithm proposed in this paper can reduce the risk of privacy leakage when a user publishes speech or text, and applies to most deep learning models and various privacy classification tasks. It is of great significance to solve the privacy protection problem for various social platforms such as mobile applications, computer software, web pages, and so on.

II. RELATED WORK

In recent years, privacy issues have received increasing attention from the Natural Language Processing (NLP) community. The work of Hovy et al. [7] has shown that textual information harbors some personal information, such as gender, age, location of the sent comment, etc. The work of Pan et al. [8] has also shown that there is a risk of privacy leakage inherent in the NLP model. Through their experimental results, Staab et al.[9] found that GPT-4[10] possesses the most superior performance in analyzing personally identifiable information.

To cope with the above problems, there have been many studies discussing defense schemes against privacy leakage attacks [3][4][11]. Elazar et al. [3] and Fernandes et al. [4] remove sensitive information from latent representations through adversarial training. Elazar and Goldberg [3] focus on the adversarial deletion of demographic attributes from textual data. They first train a classifier that accurately predicts the main task labels using some labeled data consisting of document and task labels. Then an encoder is constructed which maps to a representation vector, and an adversarial classifier is used for vector-based privacy prediction. The goal results in a trained classifier that is maximally informative about the primary task while minimizing information about the protected attributes. These works have text categorization as the target task and remove the privacy information from the intermediate representation by adversarial training. In contrast, the work in this paper does not have an adversarial model with text categorization as the target task and directly processes the raw text with word substitution to protect the sensitive information of the text authors.

Another approach is to generate new sentences with less sensitive information using privacy-aware text rewriting methods [5][6][12][13]. Xu et al. [5] designed a privacy-aware text rewriting framework based on back-translation to minimize the leakage of sensitive information. They optimized the model based on the trade-off between reconstruction loss and privacy risk loss. The reconstruction loss focuses on semantic relevance and syntactic fluency, while the privacy risk loss controls the leakage of sensitive information. The results show that

*Corresponding Author: Shuangjiao Li

the method is effective in reducing the leakage rate of sensitive information and maintaining the linguistic quality of the rewritten text. However, the idea of its privacy-aware text rewriting is based on back-translation, which translates the original text into another language and then back to the original language. In contrast, the text rewriting studied in this paper is a word-for-word replacement. [6][12] and [13] are very limited in their protection of privacy. Strengers et al. [6] can only obfuscate information about writing style. Li et al. [13] only protect against gender-specific information.

Security specifications for large language models are still focused on toxic and offensive speech [14][15]. There are fewer studies on privacy protection triggered by big models. Staab et al. [9] attempted to improve security filters for toxic speech and reduce access requests that reveal private information. However, the experimental results show that only a small fraction of the requests are rejected. The vast majority of texts containing private information cannot be recognized and filtered. The PII-Remover feature provided by Azure Language Services can help protect private information. However, only highly sensitive plaintext information is removed, and private information obtained by relying on model reasoning cannot be removed.

III. METHODOLOGY

2.1 Overall Framework

The overall framework of our network is depicted in Diagram 2. Given an input text x, it is the first text vectorized to obtain an n-dimensional vector of words $(w_1, w_2, \dots, w_n)$. The target is then randomized, i.e., the "expected label" of the target instance is randomly generated, denoted as y_{expect}, so that the attacker will not be able to predict the true label. Then take $(w_1, w_2, \dots, w_n)$ as input and generate the target instance $x^* = (w_1', w_2', \dots, w_n')$. Finally, the generated target instances are re-input to the private information prediction model to get the prediction label y_{pred}. y_{expect} should be equal to y_{pred} if the protection is successful.

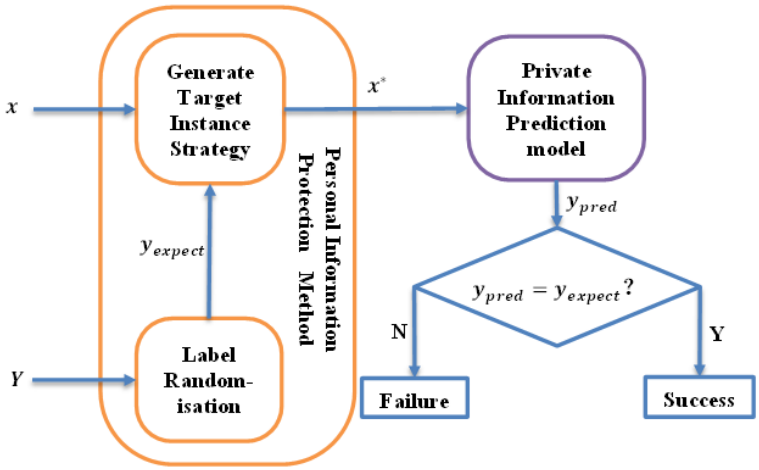


Diagram 2

The two most critical parts of the approach are the label randomization method and the generate target instance strategy.

2.2 Label Randomisation Method

The goal of our method is to tiny modify x to confuse the attacker from correctly predicting the private information. Therefore, before generating the target instance, a random function is used for a fake label y_{expect}.

$$y_{expect} = \text{randint}(0,1,2, \dots, \text{K} - 1) \rightarrow (1)$$

where K is the number of categories for privacy classification.

2.3 Generate Target Instance Strategy

The success of the defense is signified by the privacy information classifier classifying x^* to y_{expect}. There are two key issues in generating x^* by synonym substitution: the determination of the order of replacement and the selection of synonyms.

2.3.1 Replacement Order Strategy

In this paper, the concept of word saliency [11][12] is introduced to determine the order of word substitution. Word saliency is the degree of change in classification probability after masking a word. The word saliency of individual words w_i is denoted as $S(x, w_i)$, which is calculated as follows:

$$S(x, w_i) = P(y_{expect} \mid x) - P(y_{expect} \mid \hat{x}_i) \rightarrow (2)$$

where $x = w_1 w_2 \dots w_i \dots w_n$, $\hat{x}_i = w_1 w_2 \dots unknown \dots w_n$.

The word salience vector $S(x)$ for a sentence consists of all the $S(x, w)$ in x.

$$S(x) = (S(x, w_1), S(x, w_2), \dots, S(x, w_n)) \rightarrow (3)$$

2.3.2 Synonym Selection Strategy

The synonym selection strategy mainly solves the problem of choosing which word is optimal as a replacement among all the synonym candidates.

For each word w_i in x, a synonym set L_i is constructed using WordNet, which contains all the synonyms of w_i. The classification probability of y_{expect} is obtained for each synonym $w_i' \in L_i$ by the private information prediction model. Finally, the one with the largest classification probability after the replacement is selected as the replacement word w_i^* from L_i.

The specific selection formula is as follows:

$$w_i^* = arg \max_{w_i' \epsilon L_i} \{P(y_{expect} \mid w_1 w_2 \dots w_i' \dots w_n)\} \rightarrow (4)$$

After getting the replacement candidate w_i^*, replace w with w_i to get the new text x_i^*:

$$x_i^* = w_1 w_2 \dots w_i^* \dots w_n \rightarrow (5)$$

In order to comprehensively consider the influence of word salience as well as synonym quality on the defense effect, the design evaluation function $H(x, x_i^*)$ is defined as follows:

$$H(x, x_i^*) = \phi(S(x))_i \cdot P(y_{expect} \mid x_i^*) \rightarrow (6)$$

where ϕ is a softmax function.

Continuously select the synonym with the largest H for replacement to generate a new target instance x^*. Judge whether the prediction result of the classifier is equal to y_{expect}. If equal then output x^* and launch the iteration.

IV. EXPERIMENTAL RESULT

4.1 Experimental Setup

Dataset. The dataset used for this experiment was HappyDB, a corpus of over 100,000 happy moments crowdfunded through Amazon's Mechanical Turk. The dataset also provides demographic information about each interviewee, including the person's age, country, gender, marital status, and more.

Network and training details. The maximum length of the text is set to 200 words. The personal privacy information prediction model consists of a 100-dimensional word vector layer, an LSTM layer consisting of 128 cells, and a fully connected layer. Cross entropy loss function is used as a loss function. batch size is set to 32. number of training rounds is 30.

Evaluation indicators:

1) For the classification of private information, **accuracy** rate, **recall** rate, and **F1-score** are considered evaluation metrics. These three metrics can be synthesized to consider the effectiveness of the method.

2) The most intuitive metric for evaluating protection methods is the protection success rate. The **protection rate** is the percentage of successfully protected target instances in the total number of target instances. Successful protection means that the generated target instances are categorized by the Personal Privacy Information Prediction Model with labels that match the y_{expect}.

3) For the quality of the generated text, the word replacement rate and BERTScore[17] are considered evaluation metrics. The **replacement rate** is the percentage of replaced words in the original text. The lower the replacement rate is, the better the replacement effect is, and the more difficult it is for people to perceive the changes before and after the text replacement. **BERTScore** is the similarity score between the original utterance and the target instance based on BERT calculation. Its value is between -1 and 1, and the larger the value, the more similar the generated target instance is to the original text.

4.2 Protection Effectiveness

Table 1 shows the comparison of the three evaluation metrics for the original sample and the target sample generated by the protection method. The more the three evaluation metrics of the model decrease and approach 0.5 (only for the binary classification problem), the less accurate the attacker is in predicting private information about marital status and the more effective the protection method is. As can be seen from Table 1, the accuracy rate decreases by 0.229, the recall rate decreases by 0.199, the F1 score decreases by 0.185, and all

the evaluation metrics approach the random probability dramatically, indicating that the protection method proposed in this paper is effective.

Table 1: Three protection effectiveness indicators

w/wo protection	Accuracy	Recall	F1-score
w/o	0.853	0.932	0.882
w/	0.624	0.733	0.697

4.3 Target Instance Quality

Table 2 presents the results of the evaluation metrics for different instances of generating targets. As can be seen from Table 2, the protection success rate is relatively high with a very low word substitution rate. This indicates that the protection method proposed in this topic achieves the expected results by very few word substitutions. Moreover, the BERTScore is very close to 1, indicating that the semantic and syntactic features of the original samples remain largely unchanged.

Table 2: Three target instance quality indicators

Protection rate	Replacement rate	BERTScore
80.1%	7.9%	0.924

4.4 Target Instance Analysis

Table 3 shows a few examples of targets with successful protection and reversed tags compared to the original text. The bolded and red words are synonyms of the transformations that occurred before and after the protection, respectively. As can be seen from Table 3, in both examples, only a single synonym substitution reverses the original marital status label. There is no lexical or grammatical problem with the text. There is also no semantic alteration. The label inversion may even have a higher confidence interval. This shows that the algorithm proposed in this paper achieves privacy preservation through a few word substitutions. Also, the modified text still conveys the semantics of the original text and one cannot sense any change. But the attacker's natural language classifier is spoofed.

Table 3: Comparison of texts with or without protection

Instances	w/wo protection	Categories	Confidence interval	Text
Instance 1	w/o	Married	99.97%	Eating taco bell because I was starving. Spending time with my **husband**.
	w/	Unmarried	86.79%	Eating taco bell because I was starving. Spending time with my hubby.
Instance 2	w/o	Unmarried	90.26%	I found the perfect bridesmaid dresses for her **wedding**.
	w/	Married	95.78%	I found the perfect bridesmaid dresses for her marriage.

4.5 Confidence Interval Analysis

The confidence interval of the classification results of the personal privacy information prediction model can also indicate the effectiveness of the protection method. In this experiment, the confidence interval is categorized into three levels: untrustworthy: 50%~65%, more trustworthy: 65%~80%, and trustworthy: 80%~100%.

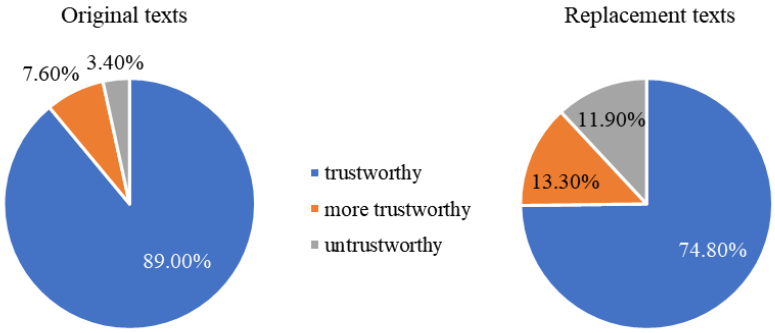


Figure 1: Percentage of confidence levels for original text and target instances

From Figure 1, it can be seen that the confidence level of the target instance after synonym replacement has been reduced, which makes it easy to understand that the prediction results of reversed labels caused by replacing only individual words are generally not as credible as the original label prediction. However, overall, the more credible and trustworthy samples still account for the vast majority, which also indicates that the protection algorithm proposed in this paper is relatively reliable.

V. CONCLUSION

In this paper, we propose a word replacement-based textual personal privacy information protection method to reduce the risk of personal privacy information leakage in text. Firstly, random target labels are generated and then the replacement strategy is decided by combining the replacement words with the maximum classification probability of the target labels and the text replacement position with maximum word saliency. Finally, the experimental results show that the protection method in this topic can effectively and reliably reduce the risk of privacy information leakage. After the specific instance comparison and target instance evaluation index also intuitively see that the modified text conforms to the grammar and the original semantics basically do not change.

REFERENCES

[1] Carlini N , Chien S , Nasr M ,et al.Membership Inference Attacks From First Principles[J]. 2021.
[2] Carlini N , Ippolito D , Jagielski M ,et al.Quantifying Memorization Across Neural Language Models[J]. 2022.
[3] Elazar Y, Goldberg Y . Adversarial Removal of Demographic Attributes from Text Data[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.
[4] Fernandes N, Dras M, Mciver A . Generalised Differential Privacy for Text Document Processing[C]// 2019.
[5] Xu Q , Qu L , Xu C , et al. Privacy-Aware Text Rewriting[C]// Proceedings of the 12th International Conference on Natural Language Generation. 2019.
[6] Strengers Y , Qu L , Xu Q , et al. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation[C]// CHI '20: CHI Conference on Human Factors in Computing Systems. 2020.
[7] D Hovy, Johannsen A , A Søgaard. User review sites as a source for large-scale sociolinguistic studies[J]. International World Wide Web Conferences Steering Committee, 2015.
[8] Pan X, Zhang M, Ji S, et al. Privacy Risks of General-Purpose Language Models[C]// To Appear in IEEE Symposium on Security and Privacy. IEEE, 2020.
[9] Staab, Robin, et al. Beyond Memorization: Violating Privacy Via Inference with Large Language Models[J]. 2023.
[10] OpenAI. Gpt-4 technical report. ArXiv, abs/2303.08774, 2023.
[11] Li Y, Baldwin T, Cohn T . Towards Robust and Privacy-preserving Text Representations[J]. 2018.
[12] Li J , Chen X , Hovy E , et al. Visualizing and Understanding Neural Models in NLP[J]. Computer Science, 2015.
[13] Li J, Monroe W, Jurafsky D . Understanding Neural Networks through Representation Erasure[J]. 2016.
[14] [1] Gehman S , Gururangan S , Sap M ,et al.RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models[J]. 2020.DOI:10.18653/v1/2020.findings-emnlp.301.
[15] Si, Waiman, et al. Why So Toxic?: Measuring and Triggering Toxic Behavior in Open-Domain Chatbots[J]. Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022.
[16] Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT[J]. 2019.