Privacy-Aware Text Rewriting

Qiongkai Xu[1,2], **Lizhen Qu**[3], **Chenchen Xu**[1,2] **and Ran Cui**[1]
[1]The Australian National University, Canberra, Australia
[2]Data61 CSIRO, Canberra, Australia
[3]Monash University, Melbourne, Australia
{Qiongkai.Xu,Chenchen.Xu,Ran.Cui}@anu.edu.au, Lizhen.Qu@monash.edu

Abstract

Biased decisions made by automatic systems have led to growing concerns in research communities. Recent work from the NLP community focuses on building systems that make fair decisions based on text. Instead of relying on unknown decision systems or human decision-makers, we argue that a better way to protect data providers is to remove the trails of sensitive information before publishing the data. In light of this, we propose a new privacy-aware text rewriting task and explore two privacy-aware back-translation methods for the task, based on adversarial training and approximate fairness risk. Our extensive experiments on three real-world datasets with varying demographical attributes show that our methods are effective in obfuscating sensitive attributes. We have also observed that the fairness risk method retains better semantics and fluency, while the adversarial training method tends to leak less sensitive information.

1 Introduction

Abuse and unauthorized use of sensitive information, such as demographic data, have become an ethical issue in our society. Such information should not be taken into account when humans or automatic decision making systems determine insurance rates, screen applicants for employment, target customers for advertising, or bank loans. Concerns about the fairness of decisions made by machine learning systems have led to an increasing body of work on the algorithmic fairness problem (Pedreshi et al., 2008; Zemel et al., 2013; Hardt et al., 2016; Chouldechova and Roth, 2018). Existing work on fairness learning largely focused on unbiased decisions based on classification. The algorithms made decisions for data consumers (e.g., bank) based on input provided by data producers (e.g., loan applicants), with the *sensitive* attributes (e.g., age, gender, and race) being exposed. Those algorithms acting as decision-makers are supposed to avoid discrimination on the basis of demographic groups of the individuals. In this case, the decision-makers are *trusted* to access sensitive attributes in a proper way.

However, we believe that it is doubtful that one can rely on algorithmic decision-makers to provide fair estimation. For example, discrimination by gender among job applicants has been reported (Calcagnini et al., 2015; Midtbøen, 2016). It was also reported that racial disparities pledged access to higher education (Farkas, 2003; Mickelson, 2003). Data producers are vulnerable to biased decisions. Therefore, we argue that data providers should also take the responsibility of protecting their own sensitive information. Although users may be allowed to conceal well-structured sensitive attributes such as age and gender, such sensitive information can still be predicted from unstructured text data (Blodgett et al., 2016; Mac Kim et al., 2017; Elazar and Goldberg, 2018; Voigt et al., 2018). As suppressing more sensitive information in text indicates more privacy, we propose a new research challenge, privacy-aware text rewriting, namely *protecting sensitive attributes in text data on behalf of data providers by rewriting the text*. A rephrased privacy-aware text should i) reduce the leakage of sensitive information; ii) retain as much semantic meaning of the original text; iii) be grammatically fluent.Compared with fair representation learning, our work focuses on text in string form.

Transforming text into a form with less sensitive information is challenging in two ways. The first challenge is that there is a trade-off between privacy preservation and semantic relevance or fluency during rewriting. For example, *"I am a software engineer with 18 years of working experience."* shows that the author is probably over 40

Proceedings of The 12th International Conference on Natural Language Generation, pages 247–257,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

years old. Replacing '18 years' with 'more than 10 years' altogether reduces the leakage of age information with slight shift of its semantic meaning. Removing '18 years of working experience' provides stronger privacy protection, while the semantic loss is greater at the same time. The data providers should leverage the trade-off depending on varying scenarios. Another challenge is that the indicators of such sensitive attributes are subtle. For example, *"I went to the restaurant with my boyfriend. The food is yummy!"* is a post from social media. *'boyfriend'* is an explicit indicator for female user, while *'yummy'* is an implicit indicator which can be ignored by humans and captured by machine learning models. Automatic text rewriting tools help people detect and modify the subtle indicators in their text.

To address the aforementioned problems, we propose to develop a tool that rewrites text into less sensitive ones. In this work, we design a privacy-aware text rewriting framework based on back-translation to reduce the leakage of sensitive information. The models are optimized according to the trade-off between a reconstruction loss and a privacy risk loss. The reconstruction loss focuses on semantic relatedness and grammatical fluency, and the privacy risk loss controls the leakage of sensitive information. We further explore two variants of the approach. The first method formulates the privacy risk as an adversarial loss derived from a text classifier. The second method derives an upper bound of an approximate fairness risk measurement on text data, which minimizes the discrepancy of generated text among different demographic groups. Finally, we conduct extensive experiments on three datasets with varying demographic groups (i.e. Politics, Gender, and Race). The results demonstrate the effectiveness of our methods in terms of reducing the leakage rates of sensitive information and retaining linguistic quality of the rewritten text. This work provides a novel framework for systematic research on privacy-aware text rewriting, including datasets, evaluation metrics and rewriting methods, which will promote the interest in privacy preservation in our research community.

The main contributions of this work are:

- To provide the first proposal for protecting sensitive attributes in text on behalf of data providers.
- To design a privacy-aware back-translation method for protecting sensitive information in rewritten text.
- To provide datasets and evaluation metrics for appropriate validation of method effectiveness.

2 Privacy-Aware Text Rewriting

Privacy-aware rewriting modifies text to obfuscate a sensitive attribute. The bespoke methodologies aim to minimize the loss of fluency as well as the change in the underlying semantics. We consider a setup in which we have a set of input text $\{X_1, ..., X_N\}$, where each text X_i is a word sequence $\langle x_1, ..., x_l \rangle$. Each text is associated with a sensitive attribute S, such as gender or race. The goal is to find a privacy-aware translator $f(X) : X \rightarrow Y$ to modify X into another word sequence $Y = \langle y_1, ..., y_l \rangle$, such that an attacker $g(Y) : Y \rightarrow S$ fails to predict the values of the sensitive attribute S from the translated text Y.

2.1 Privacy-aware Back-Translation Model

Privacy-aware rewriting can be regarded as a special monolingual machine translation (MT) task, which aims to remove sensitive information through rephrasing. In our experiment, there is no existing parallel corpus to learn the patterns of privacy-preserved rewriting. We use Back-Translation to obtain a meaning-preserving representation in the target language, and translate the sentences back to the source language (Prabhumoye et al., 2018). Since we aim to preserve sensitive information, we consider the risk from an attacker in the back-translation phase.

In our work, the source language is English and the target language is French. Let Z denote the space of target language, we build two translation models $\mathcal{T}_{\text{en}\rightarrow\text{fr}} : X \rightarrow Z$ and $\mathcal{T}_{\text{fr}\rightarrow\text{en}} : Z \rightarrow X$, respectively. We use the Transformer-based model (Vaswani et al., 2017) for each translation model. The back-translation procedure is formulated as,

$$f(X) = \mathcal{T}_{fr\rightarrow en}(\mathcal{T}_{en\rightarrow fr}(X)) \quad (1)$$

For each input text, the outcome of this model is a sequence of words in English.

The goal of learning privacy-aware back-translation is two-fold. Firstly, it aims to find an optimal predictor f^* that minimizes an expected reconstruction loss $\mathbb{E}_{X,Y}[\mathcal{L}(f(X), Y)]$

with $\mathcal{L}(f(X), Y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, which measures the discrepancy between predicted sequences $f(X)$ and true target sequences Y. Secondly, the predictor should be reasonably fair to S by achieving a low risk loss with regard to privacy $\mathcal{R}(X, Y, S) : \mathcal{X} \times \mathcal{Y} \times \mathcal{S} \rightarrow \mathbb{R}$. Let $\mathcal{F}$ denote the space of all possible predictors, we find the optimal rewriting model f^* by

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\mathcal{L}(f(X), Y)] + \alpha \mathcal{R}(X, Y, S) \tag{2}$$

where α controls the degree of privacy protection.

2.2 Adversarial Classifier

Given an accurate classifier, the risk of privacy is able to be estimated by the negative classification loss on the sensitive information. Our target is finding the representations that are good at reconstructing the sentences, while poor in predicting sensitive labels. The setting is well-aligned with generative adversarial networks (Goodfellow et al., 2014). We construct the back-translation model as $f(X) = m(h(X))$, where $h(X)$ employs the two translators to map X into a sequence of hidden representations of decoded words in the source language. Then, $m(\cdot)$ maps the hidden representations into the corresponding words. An adversarial classifier $adv(h(X))$ is a linear classifier, which takes the mean of all hidden representations from $h(X)$ to predict S. The risk is formulated as adversarial classification loss $\mathcal{L}_c(adv(h(X)), S)$. The encoder $h(\cdot)$ is trained to fool the adversarial classifier $adv(\cdot)$ while optimizing the back-translation predication $f(X)$ in Eq.(4). Eq.(3) merely optimizes the adversarial classifier. The training is conducted by jointly optimizing the following two objectives:

$$\arg\min_{adv} \mathcal{L}_c(adv(h(X)), S) \tag{3}$$

$$\arg\min_{h,m} \mathcal{L}_g(m(h(X)), X) - \alpha \mathcal{L}_c(adv(h(X)), S) \tag{4}$$

where $\mathcal{L}_g$ is the cross entropy loss with Label Smoothing (Szegedy et al., 2016) for the transformer-based generator and L_c is the cross entropy loss for the adversarial classifier. The negative parameter $-\alpha$ is implemented by a gradient-reversal layer (GRL)(Ganin and Lempitsky, 2015) during back-propagation and α controls the intensity of adversarial training.

2.3 Fairness Risk Measurement

In this section, we define the privacy risk loss using fairness risk measurement. The perfect fairness for rewriting is a statement of conditional independence of generated text $Y \perp\!\!\!\perp S|X$. Holding such condition, the sensitive translator conduct similar generation results. Therefore, attackers will not be able to infer the dependent attributes. A *privacy-aware* translator $f(X)$ learns a distribution $P(Y|X)$, while $P(Y|X, S = a)$ denotes the distribution of a *subgroup* translator depending on a particular demographic group attribute S. The conditional independence is formulated as,

$$P(Y|X) = P(Y|X, S = a) \tag{5}$$

Agarwal et al. (2018) pointed out that given finite samples in training data, it is impossible to ensure perfect fairness on the test sample. An approximate formalism of fairness measurement is used to quantify the discrepancy of demographic parities, namely maximal deviation between subgroup predictions (MDSP) (Calmon et al., 2017).

$$\sup_{y,s,s'} |Pr(\hat{Y} = y|S = s) - Pr(\hat{Y} = y|S = s')| \tag{6}$$

where $\hat{Y}$ is a single variable.

Inspired by the single-variable MDSP, we define the sequential MDSP (SMDSP) for text rewriting as,

$$\sup_{a \in S} |\log P(Y|X) - \log P(Y|X, S = a)| \tag{7}$$

where Y is the generated sequences. We obfuscate the sensitive attribute by reducing the discrepancy between privacy-aware translator and the most different subgroup translator.

The challenge of using the SMDSP is that it is optimized on the whole sequences. However, the state-of-the-art encoder-decoder architecture (Vaswani et al., 2017; Klein et al., 2017) generate words in a word-by-word manner.We derive an upper bound of SMDSP by applying calculus on the sequential deviation

$$\begin{aligned}
&D(X, Y, S = a) \doteq |\log P(Y|X) - \log P(Y|X, S = a)| \\
&= |\sum_{i=1}^{l} \log P(y_i|X, y_{<i}) - \sum_{i=1}^{l} \log P(y_i|X, y_{<i}, S = a)| \\
&\leq \sum_{i=1}^{l} |\log P(y_i|X, y_{<i}) - \log P(y_i|X, y_{<i}, S = a)| \\
&\doteq \mathcal{U}_a(X, Y)
\end{aligned}$$

The composition of MDSP for each word is an upper bound of SMDSP.

$$\mathcal{R}_u(X, Y, S) = \sup_{a \in S} \mathcal{U}_a(X, Y) \quad (8)$$

We replace the approximate fairness risk by its upper bound Eq.(8) and obtain a joint training objective.

$$\mathcal{L}_\alpha(X) = \mathcal{L}(f(X), X) + \alpha \mathcal{R}_u(X, Y, S) \quad (9)$$

In training, each subgroup translator is pre-trained beforehand with the training data labeled with the corresponding sensitive attribute value. Their parameters are kept fixed when minimizing the privacy-aware rewriting model.

3 Experimental Setup

3.1 Datasets

In this paper we conduct experiments on three tasks, which can lead to potential social-good applications, namely obfuscating gender, political slant and race of the authors. .

Gender (Reddy and Knight, 2016) is a dataset of reviews from Yelp annotated with the gender of the authors, either male or female. The sentences with low indication of gender (likelihood of gender lower than 0.7) is filtered out.

Politics (Voigt et al., 2018) is a dataset of comments on Facebook posts from 412 members from the United States Senate and House. Each comment is associated with the corresponding Congressperson's party affiliation as the sensitive attribute, $S \in \{\text{democratic}, \text{republican}\}$.

Race (Blodgett et al., 2016) is a dataset based on the dialectal tweets corpus (DIAL), including 59.2 million tweets. The tweets are categorized into African-American English (AAE) or Standard American English (SAE), which is highly correlated to the race of the author. The predictor takes into account both the content of the tweets and the geolocations of the the authors. We filter out the samples with predicted confidence lower than 80%, and tweets with less than 3 words. We consider race as sensitive information of the dataset. We also maintain the sentiment classification as a target task for this corpus to check if the sentiment information is still preserved after rewriting. The sentiment labels are derived from emojis which are associated with sentiments.

All the aforementioned corpora are split into four disjoint parts: **Class**, training corpus for sensitive attribute classifier; **Train**, training corpus for privacy-aware text rewriting; **Valid**, validation set; and **Test**, test set. The number of sentences for each split of these datasets are listed in Table 1. The datasets are publicly available at `https://github.com/xuqiongkai/PATR`

Dataset	Class	Train	Valid	Test
Gender	2.6M	200K	4K	4K
Politics	80K	200K	4K	4K
Race	80K	100K	4K	4K

Table 1: Data splits of `Gender`, `Politics` and `Race`.

3.2 Models

We consider the following three models for privacy-aware text rewriting. **Back Trans** is the back translation model considered as baseline. **Adv** is the model using adversarial training. **SMDSP** model use Sequential Maximal Deviation between Subgroup Predictions. We also compare the quality of generated text of our systems with those of an open-domain **Paraphrase** generation system (Iyyer et al., 2018).

3.3 Implementation Details

We use Transformer (Vaswani et al., 2017) as the translation architecture in our experiments. We re-implement the transformer model based on OpenNMT (Klein et al., 2017). In our experiments, we use the same configurations, including 2 encoder and decoder layers, 256-dimensional word embedding and 256-dimensional hidden layers, drop out rate 0.1, label smoothing weight 0.1. All models use Beam Search decoding algorithm with beam size 5.

We train English-French machine translation (En-Fr) and French-English back-translation (Fr-En) using Europarl v7 from WMT15 (Bojar et al., 2015). The words are tokenized using Moses tokenizer (Koehn et al., 2007). Our translation system achieves the BLEU scores of 36.24% and 37.36% on En-Fr and Fr-En, respectively. The En-Fr model is used to generate the parallel corpus for all experiments.

3.4 Evaluation

The generated sentences are evaluated according to both linguistic quality of the sentences and obfuscation of the sensitive attribute. For each of these two aspects, we conduct automatic evaluation and human evaluation, respectively.

Linguistic Quality focuses on evaluating the quality of the results based on their semantic relevance to the original text and gramatical fluency of the generated sentences. We adopt four automatic evaluation metrics, BLEU, GLEU, METEOR and WMD. BLEU (Papineni et al., 2002) and GLEU (Wu et al., 2016) measure the n-gram matching between hypothesis and reference, where GLEU considers both precision and recall. METEOR (Banerjee and Lavie, 2005) further applies stemming and synonym matching. Word Mover Distance (WMD) (Kusner et al., 2015) calculates the optimal transport distance between word embedding in original and generated sentences[1]. Intuitively, BLEU and GLEU evaluate fluency of the sentence as they are based on the quality of n-grams, while WMD measures semantic relevance as words can be regarded as atom semantic components of sentences.

We also conduct human evaluation to judge the fluency and relevance of the results[2]. For each set of the results, two annotators are asked to judge the quality of the results between the scales of 1-5. The Kappa coefficients (McHugh, 2012) on `Gender`, `Politics` and `Race` are 0.45, 0.47 and 0.74, respectively.

Obfuscation evaluates the leakage of sensitive attributes of generated text. For automatic evaluation, we estimate the probability of sensitive attribute on generated sentences using a Logistic Regression with L2 regularization (Pedregosa et al., 2011). For all the experiments, we use top 3K frequent words as features. Based on the prediction of classifier $p_i = P(S = i|X)$, we propose to evaluate the obfuscation of the results using the following three metrics:

1. **Entropy** evaluates the averaged entropy ($\sum_i -p_i \log p_i$) of all predictions. Higher Entropy indicates better less sensitive information leakage.

2. **P-Acc**, prediction accuracy, calculates the portion of correct prediction of the sensitive attribute. In the case of binary classification, the score is better if it is closer to 50%.

3. **M-Acc**, modification accuracy, calculates the label probabilities of source and generated sentences. If the probability of the sensitive attribute decreases after rewriting, the modification is accepted. M-Acc counts the rate of accepted sentence modifications.

In human evaluation, annotators are asked to judge the sensitive attribute values of 300 sampled sentences in test set. We use accuracy to evaluation the awareness of sensitive information by human and automatic annotators. Due to the fact that human judgments underperform automatic judgments (see Table 4), we rely more on automatic metric to evaluate the rewriting results.

4 Results and Analysis

We first conduct human evaluation and discuss their relation to automatic evaluation metrics with regard to semantic relevance, grammatical fluency and obfuscation. Then, we compare our privacy-aware models according to linguistic quality and obfuscation. Later on, we test the semantic loss of our models on the target task. Finally, we provide some sample outputs for case study.

4.1 Human Evaluation

Firstly, we ask human annotators to evaluate linguistic quality of Back Trans, Adv ($\alpha = 1$) and SMDSP ($\alpha = 1$), based on the rewriting results from 300 test samples, with regard to fluency (Flu) and relevance (Rel)[3]. We calculate the Pearson Correlation between human and automatic evaluation metrics. Table 2 shows the correlation of semantic relevance between human and automatic evaluation. WMD is clear winner among all automatic metrics across the three datasets. According to Table 3, GLEU is the measure that most correlated to human judgement in terms of fluency, though METEOR falls slight short on the gender corpus. Unsurprisingly, the widely used BLEU is the relatively less correlated to human perception, which was also observed in machine translation (Wu et al., 2016; Callison-Burch et al., 2006).

Secondly, we compare the performance of predicting sensitive information between human annotators and automatic classifiers. We ask human annotators to classify the sensitive attributes of 300 original sentences in test set. The accuracy of the annotations are illustrated in Table 4.

[1]We use pre-trained word2vec model trained on Google News dataset from `https://code.google.com/archive/p/word2vec/`.

[2]We refer readers to Appendix A for more details about the annotation guideline.

[3]We refer readers to Appendix A with more details on annotation guideline.

Exp	BLEU	GLEU	METEOR	WMD
Gender(Adv)	0.489	0.557	0.559	**0.651**
Gender(SMDSP)	0.414	0.507	0.511	**0.645**
Politics(Adv)	0.372	0.460	0.496	**0.573**
Politics(SMDSP)	0.358	0.474	0.476	**0.563**
Race(Adv)	0.311	**0.545**	0.532	0.127
Race(SMDSP)	0.242	0.386	0.367	**0.382**

Table 2: Correlation between semantic relevance and automatic evaluation metrics on `Gender`, `Politics` and `Race`. The most correlated automatic metrics are **bold**.

Exp	BLEU	GLEU	METEOR	WMD
Gender(Adv)	0.265	**0.287**	0.222	**0.297**
Gender(SMDSP)	0.192	**0.231**	0.186	**0.361**
Politics(Adv)	0.180	**0.260**	**0.277**	0.200
Politics(SMDSP)	0.149	**0.236**	**0.236**	0.231
Race(Adv)	0.168	**0.403**	**0.433**	0.333
Race(SMDSP)	0.068	**0.150**	**0.124**	0.046

Table 3: Correlation between fluency and automatic evaluation metrics on `Gender`, `Politics` and `Race`. Top two correlated automatic metrics are **bold**.

To our surprise, human judgments are more than 10% worse than our classifiers on all the experiments. For `Politics`, we ask one more annotator for additional annotation and the accuracy of the annotation is still lower than 65%. After investigating the datasets, we found that a large proportion of samples are difficult for human annotators while our classifier can predict them correctly. For example, in `Gender`, human struggled in deciding whether "the food is delicious" and "the people were nice" are posted by male or female authors. For `Politics`, we observe several cases that human tends to annotate them with the opposite political slant when the sentences are in negative sentiment, while actually the speaker and the mentioned people support the same party, e.g., "Patty Murray couldn't be any more dishonest than this!". Other examples like "today is such a wonderful day!" and "God bless you guys" are neutral to our annotators. Correctly annotating these samples might require extensive background in American politics[4]. To sum up, human annotators fail to incorporate subtle indicators into their decision, however, the classifiers manage to detect them.

The human evaluation studies conclude that i) we can rely on sensitive attribute classifiers for obfuscation evaluation, and ii) we should look at WMD for semantic relevance and GLEU for fluency.

	Gender	Politic	Race
Automatic	**77.3**	**93.7**	**82.7**
Human	66.0	60.3	71.0

Table 4: Comparison of human and automatic judgments on `Gender`, `Politics` and `Race`.

4.2 Adversarial Learning vs. SMDSP

We conduct automatic evaluation on text generated by Back Trans, Adv and SMDSP. The overall observations are i) Back Trans provides a preliminary baseline for our task; ii) both Adv and SMDSP are able to reduce the leakage of sensitive information; and iii) SMDSP retains better linguistic quality, while Adv manages to preserve sensitive information.

We first compare the linguistic quality of the results in Table 5. The Back Trans outperforms both Adv and SMDSP on average because it does not cope with sensitive attributes in training. The performance of Adv model with the highest α obtains less than half GLEU than that of Back Trans. Although SMDSP with higher α also shows performance reduction, the quality of generated text are still competitive with Back Trans, with less than 10% score reduction. In particular, SMDSP with ($\alpha = 1$) achieves even higher GLEU on both `Politics` and `Race` than the baseline. We attribute this to the regularization effect of SMDSP on language modeling. Results of human evaluation are coherent to automatic evaluation, in Table 7. SMDSP achieves highest fluency results and competitive relevance results.

Then, we show the obfuscation performance in Table 6. Back Trans is a competitive baseline that obfuscates the classifiers to some extent. Adv and SMDSP are able to further reduce the obfuscation score on all three datasets. Generally, models with higher α achieve better obfuscation performance. Adv tend to be more aggressive on privacy preservation than SMDSP. However, we observe that Adv acquires better privacy preservation by sacrificing the linguistic quality, e.g., Adv ($\alpha = 5$) basically chooses to 'keep silent' (produces almost no words) to protect the sensitive information on `Politics`[5]. We believe that generating totally non-sense sentences is too conservative for our task. On the other hand, SMDSP manages to protect sensitive attribute while keeping the semantic meaning as much as possible. For

[4]The top weighted words of male or female for `Gender`, democratic or republican for `Politics`, and SAE for `Race` are listed in Appendix B to show the difficulty for human annotators to capture subtle indicators.

[5]All the generated sentences are empty on test set.

Model	Gender			Politics			Race		
	GLEU	METEOR	WMD	GLEU	METEOR	WMD	GLEU	METEOR	WMD
Back Trans	**45.14**	**37.16**	**1.012**	37.29	**36.78**	**1.039**	23.09	26.94	1.460
Adv($\alpha = 1$)	44.11	36.76	1.023	29.44	33.55	1.125	12.94	18.07	**1.303**
Adv($\alpha = 2$)	40.29	34.34	1.117	23.20	26.82	1.261	12.75	18.39	1.430
Adv($\alpha = 5$)	22.98	23.32	1.561	N/A	N/A	N/A	9.67	17.03	2.242
SMDSP($\alpha = 1$)	44.17	36.69	1.031	**38.43**	36.59	1.044	**24.77**	**28.15**	1.483
SMDSP($\alpha = 2$)	43.10	35.84	1.062	38.01	36.36	1.056	23.95	27.49	1.501
SMDSP($\alpha = 10$)	41.54	35.09	1.101	36.40	35.96	1.069	23.10	26.99	1.531
SMDSP($\alpha = 100$)	40.90	34.64	1.122	36.84	35.64	1.082	22.74	26.81	2.242

Table 5: Automatic evaluation of linguistic quality on Gender, Politics and Race.

Model	Gender			Politics			Race		
	Entropy	P-Acc	M-Acc	Entropy	P-Acc	M-Acc	Entropy	P-Acc	M-Acc
Test(Ori)	0.5544	77.45	-	0.4873	93.05	-	0.3586	86.33	-
Back Trans	0.5617	72.45	48.90	0.5011	85.55	56.03	0.3960	74.68	62.35
Adv($\alpha = 1$)	0.5649	72.50	49.58	0.5026	84.90	57.25	0.4386	74.08	66.80
Adv($\alpha = 2$)	0.5644	70.23	52.73	0.5542	73.60	68.65	0.4623	73.40	69.13
Adv($\alpha = 5$)	0.5754	66.80	59.78	0.6931	50.00	93.15	0.5268	65.75	73.58
SMDSP($\alpha = 1$)	0.5711	71.80	50.18	0.5059	85.20	57.33	0.3989	74.85	62.48
SMDSP($\alpha = 2$)	0.5759	71.08	52.15	0.5066	84.95	58.35	0.4013	74.40	63.40
SMDSP($\alpha = 10$)	0.5768	70.88	53.05	0.5089	85.13	59.23	0.4007	74.08	63.65
SMDSP($\alpha = 100$)	0.5803	70.73	54.78	0.5129	85.08	59.90	0.4069	74.10	64.80

Table 6: Automatic evaluation of Obfuscation on Gender, Politics and Race.

Model	Gender		Politics		Race	
	Flu	Rel	Flu	Rel	Flu	Rel
Back Trans	4.68	4.09	**4.60**	**4.31**	4.31	3.88
Adv	4.66	4.13	4.42	4.01	3.84	3.53
SMDSP	**4.73**	**4.14**	**4.60**	4.21	**4.37**	**3.98**

Table 7: Human evaluation of fluency (Flu) and relevance (Rel) on Gender, Politics and Race based on the results of Back Trans, Adv ($\alpha = 1$) and SMDSP ($\alpha = 1$) with the scales of 1-5.

example, SMDSP ($\alpha = 1$) achieves both higher relevance score and better obfuscation score than Adv ($\alpha = 1$) on Gender and Politics.

Finally, we demonstrate the training stability of our models. The reconstruction losses of each model on validation set of Gender, Politics and Race are shown in Figure 1. We pre-train the back translation model for 10 epochs on Gender and 20 epochs on Politics and Race. Then, we train Adv model and SMDSP model based on the pre-trained model. We also include the pre-trained model with the same total number of training epochs in Black lines. After pre-training, Back-Trans models start to overfit and get slightly worse results on validation set. In most cases, the losses of Adv are higher than Transformer, and higher adversarial training intensity α decreases the performance of translation model. Adv ($\alpha = 5$) is not included in the plots, because their losses are out of the range. In contrast, SMDSP achieves better performance than Adv. The performance of SMDSP is even better than Back Trans on Gender and Race.

4.3 Target Task Performance

We evaluate sentiment classification (Sent) as the target task and racial (Race) as sensitive attribute on the Race. As shown in Table 8, the prediction performance of both Race and Sent using Adv models decrease as the hyperparameter α increases. Such trend shows that Adv improves privacy preservation by obfuscating the semantic meaning of the original text. In contrast, Risk models successfully decrease the accuracy on Race, while preserving the accuracy on Sent, showing the robustness of the model on preserving semantic meanings of the text.

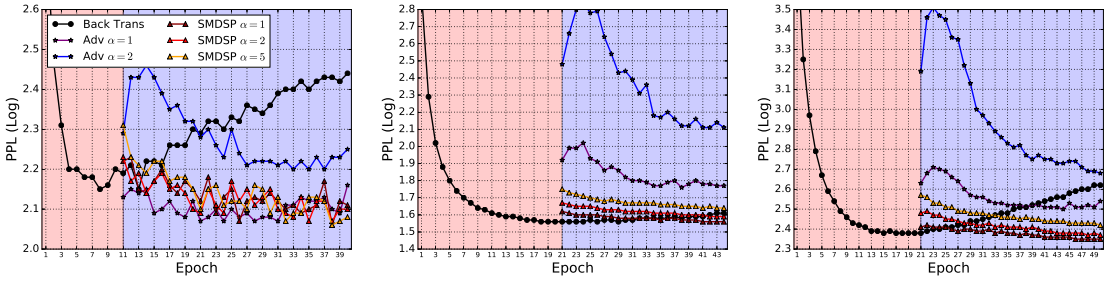
4.4 Case Study

We demonstrate generated examples in Figure 2[6]. For Gender, Back Trans generates the words with clear tendency of gender, such as 'yummy' and 'girlfriend', while privacy-aware models use 'delicious', 'amazing' and 'friend' instead. For Politics, Adv and SMDSP skip the name after Sir to hide the political affiliation of the person. In the second example, Adv and SMDSP replace 'love you' with 'help' to reduce the political slant.

5 Related Work

Achieving fairness or preserving privacy through removing sensitive information from text has been explored by adversarial training (Li et al., 2018;

[6]Because the samples in Race are full of porny and violent words, they are excluded in the paper.



(a) Gender (b) Politics (c) Race

Figure 1: Log perplexity(PPL) on valid set of `Gender`, `Politics` and `Race`. Red areas indicate pre-training epochs and Blue areas represent the epochs for privacy-aware training.

Dataset	Original	Back Trans	Adv	SMDSP
Gender	food is always delicious ! (Female)	the food is always yummy !	the food is always delicious !	the food is always amazing !
Gender	I went with my girlfriend , and another couple . (Male)	I went with my girlfriend , and another couple .	I went with my friend , and another couple .	I went with my friend , and another couple .
Politics	Sir Scott , a limited attack will make any solution for Syria nearly impossible . (Republican)	Sir Scott , a limited attack will make almost impossible to Syria .	Sir , a limited attack will almost impossible attack Syria .	Sir , a limited attack will make almost impossible for Syria .
Politics	love you U.S. senator al franken (Democratic)	love you U.S. senator al franken	help U.S. senator al franken	help U.S. senator al franken

Figure 2: Sample of original text, with sensitive attribute labels, and corresponding rewritten text using Back Trans, Adv ($\alpha = 1$) and SMDSP ($\alpha = 1$) on `Gender` and `Politics`.

Model	Race	Sent
Test(Ori)	86.33	74.08
Back Trans	74.68	70.18
Adv($\alpha = 1$)	74.08	70.15
Adv($\alpha = 2$)	73.40	69.88
Adv($\alpha = 5$)	65.75	65.70
SMDSP($\alpha = 1$)	74.85†	69.88
SMDSP($\alpha = 2$)	74.40	70.23†
SMDSP($\alpha = 5$)	74.30	70.15†
SMDSP($\alpha = 10$)	74.08	70.60
SMDSP($\alpha = 100$)	74.10	70.83†

Table 8: Prediction accuracy (P-Acc) of classification results of race and sentiment classification task on `Race`. The results with higher accuracy than Back Trans are marked with daggers (†).

Elazar and Goldberg, 2018; Coavoux et al., 2018) and differential privacy (Fernandes et al., 2018). These work considers text classification as the target task and avoid data leakage by learning privacy-preserving latent representations. In contrast, our work aims to generate text in string form to protect sensitive information for data producers, which can be viewed as a special form of fair representation learning.

Paraphrase generation and text simplification are two tasks closely related to privacy-aware rewriting. Most models are based on monolingual machine translation (Ibrahim et al., 2003; Zhao et al., 2010; Wubben et al., 2012; Xu et al., 2012, 2016; Nisioi et al., 2017; Wang et al., 2016). Our work focuses on generating obfuscated text in order to conceal sensitive attribute.

There is a fast growing body of work on stylistic language generation, which focus on generating text with particular styles (e.g., humour or romantic) while trying to retain the meaning of text (Mathews et al., 2016; Fu et al., 2018; Su et al., 2018; Xu et al., 2019). Style transfer is also considered as text rewriting, which adds style information to text (Shen et al., 2017; Prabhumoye et al., 2018). In contrary, our work tries to eliminate the additional sensitive information.

6 Conclusion

In order to protect sensitive information in text, we propose a privacy-aware back-translation method for text rewriting. Adversarial training and fairness risk measurement based approaches are proposed to incorporate the privacy risk. We propose the evaluation metrics for the task to assess semantic relevance, fluency and obfuscation of the results. Our experimental results show that both methods reduce the leakage of sensitive information, and the fairness risk based method is able to better retain fluency and relevance than the adversarial one.

Acknowledgement

We gratefully acknowledge Philip Cohen for his insightful advice and encouragement on this project, as well as Alasdair Tran and Dawei Chen for their suggestions on the paper.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 2015 Workshop on Statistical Machine Translation*.

Giorgio Calcagnini, Germana Giombini, and Elisa Lenti. 2015. Gender differences in bank loan access: an empirical analysis. *Italian Economic Journal*, 1(2):193–217.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of Advances in Neural Information Processing Systems*, pages 3992–4001.

Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

George Farkas. 2003. Racial disparities and discrimination in education: What do we know, how do we know it, and what do we need to know? *Teachers College Record*, 105(6):1119–1146.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2018. Generalised differential privacy for text document processing. *arXiv preprint arXiv:1811.10256*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of International Conference on Machine Learning*, pages 1180–1189.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Proceedings of Advances in neural information processing systems*, pages 3315–3323.

Ali Ibrahim, Boris Katz, and Julie Qiaojin Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Demo and Poster sessions*, pages 177–180.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*.

Sunghwan Mac Kim, Qiongkai Xu, Lizhen Qu, Stephen Wan, and Cécile Paris. 2017. Demographic inference on twitter using recursive neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 471–477.

Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3574–3580.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Roslyn Arlin Mickelson. 2003. When are racial disparities in education the result of racial discrimination? a social science perspective. *Teachers College Record*.

Arnfinn H Midtbøen. 2016. Discrimination of the second generation: Evidence from a field experiment in norway. *Journal of International Migration and Integration*, 17(1):253–272.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 866–876.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6830–6841.

Jinyue Su, Jiacheng Xu, Xipeng Qiu, and Xuanjing Huang. 2018. Incorporating discriminator in sentence generation: a gibbs sampling method. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pages 5998–6008.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of AAAI*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Googles neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Qiongkai Xu, Chenchen Xu, and Lizhen Qu. 2019. Alter: Auxiliary text rewriting tool for natural language generation. *arXiv preprint arXiv:1909.06564*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Alan Ritter, William B. Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333.

Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. 2010. Leveraging multiple mt engines for paraphrase generation. In *Proceedings of COLING*.