Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

利用差分隐私噪声扰动的数据脱敏方法

硕士研究生 关业礼

2021 年 12月 05日

内容提要

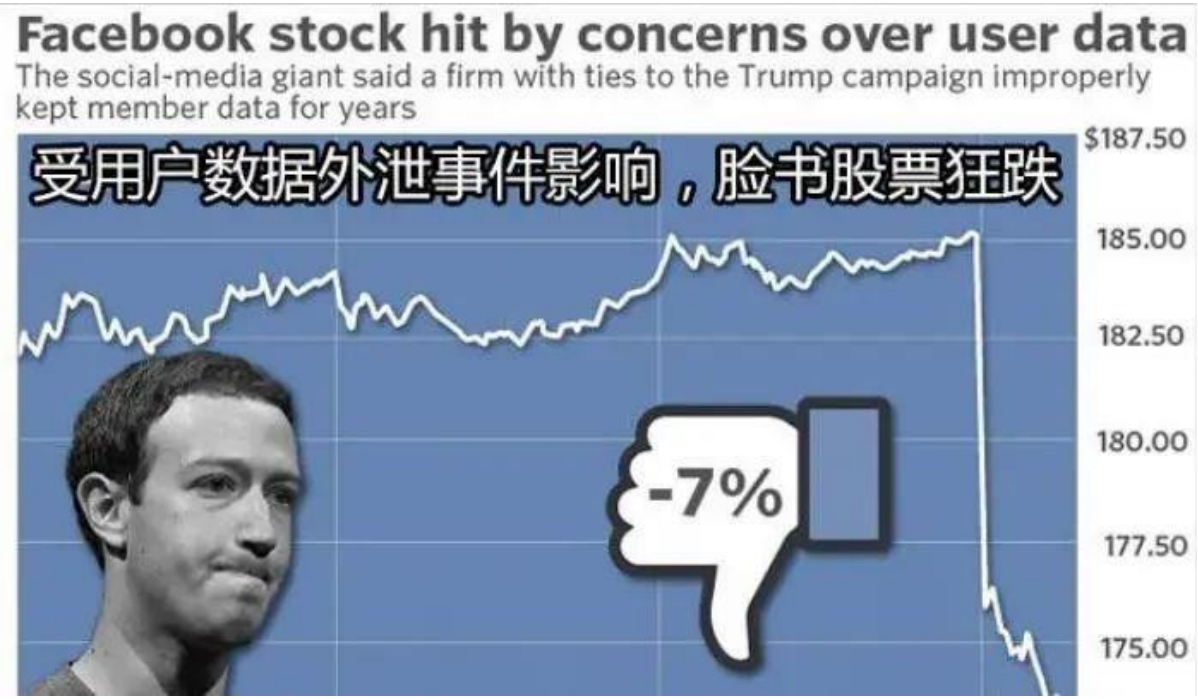
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

内容提要

- 预期收获
  - 1.了解数据脱敏的基本概念。
  - 2.理解差分隐私噪声扰动的技术原理。
  - 3.了解差分隐私在数据脱敏上的应用及研究方向。

背景介绍

- 2018年，美国社交媒体脸书（Facebook）的5000万用户信息在用户不知情的情况下，被政治数据公司“剑桥分析”获取并利用。
- 包含敏感信息的用户信息能起到什么作用?
  - 精准投放广告或咨询内容;
  - 恶意人肉搜索;
  - 电信诈骗。

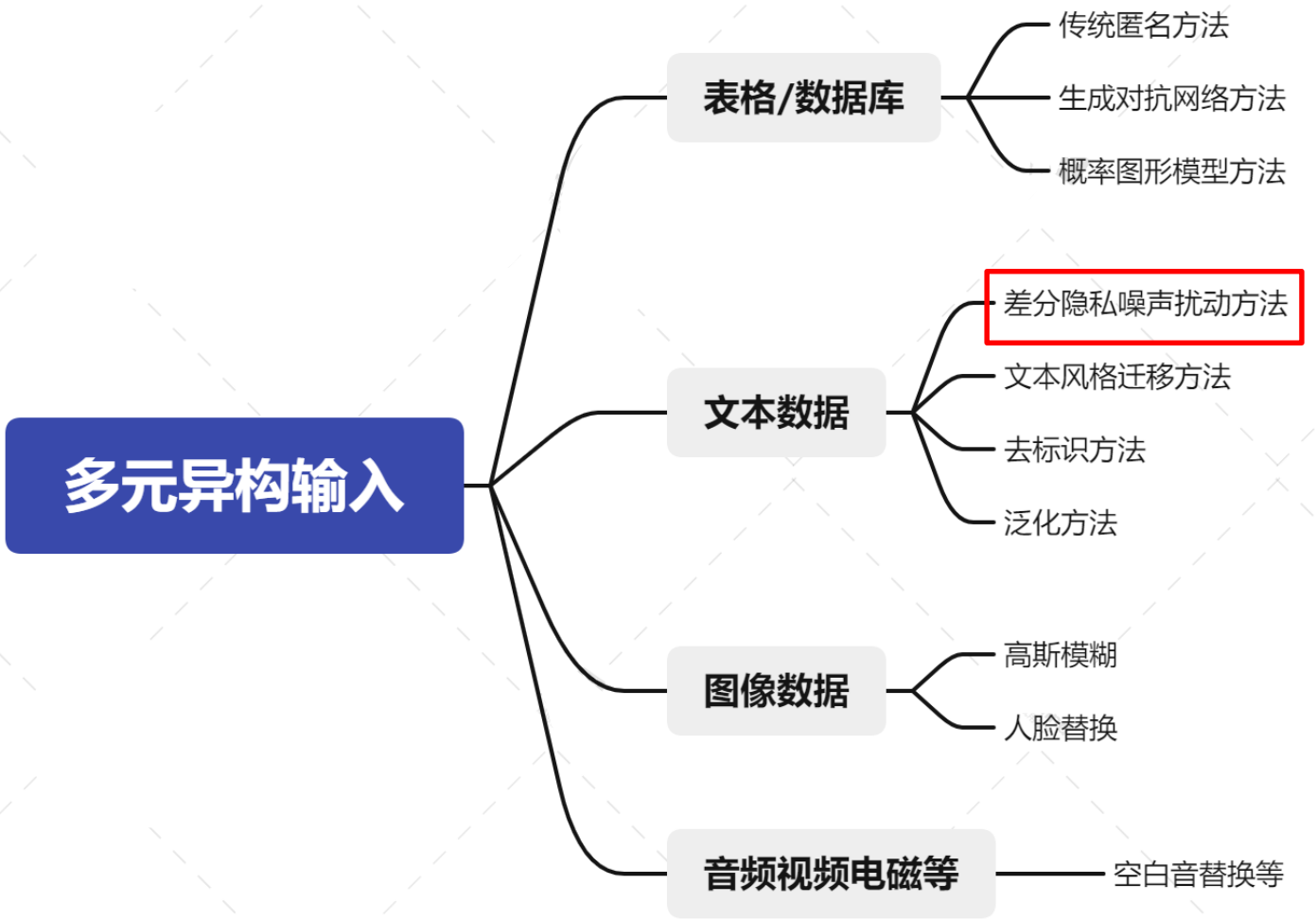


背景介绍

- 数据脱敏是指敏感数据的漂白、变形或去隐私化，从而有效可靠地保护敏感数据。数据脱敏应满足以下要求：
  - 脱敏数据尽可能保留原始数据特征，包括分布曲线和统计特征；
  - 避免从脱敏数据中恢复完整的原始数据；
  - 脱敏后的数据能够开发、生产和发布，而不泄露敏感信息。

背景介绍

- 不同输入下数据脱敏所使用的方法：



基本概念

基本概念

- **差分隐私定义：**
  - **对于一对相邻的数据集D和D’，查询获得相同结果的概率非常接近。ε-差分隐私的定义公式如下：**

含义：无论一个个体数据是否在数据库中，攻击者通过查询获得近乎相同的数据。

随机化算法M是在D上做任意查询操作，查询后对数据结果添加噪声。

两个数据库加上统一随机噪声之后查询结果为c的概率。

$$e^{-\varepsilon} \leq \frac{\Pr(M(D) = c)}{\Pr(M(D') = c)} \leq e^{\varepsilon}$$

隐私预算，反映差分隐私保护程度。

D和D’是相邻数据集。

姓名	手机	是否单身
张三	130…	是
…	…	否

姓名	手机	是否单身
张三	130…	是
李四	150…	是
…	…	否

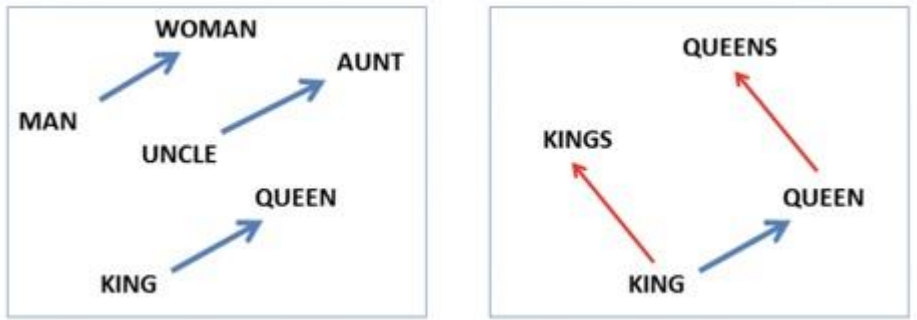
基本概念

- 词嵌入向量
  - 单词或词组映射到多维实数空间中形成词嵌入向量。

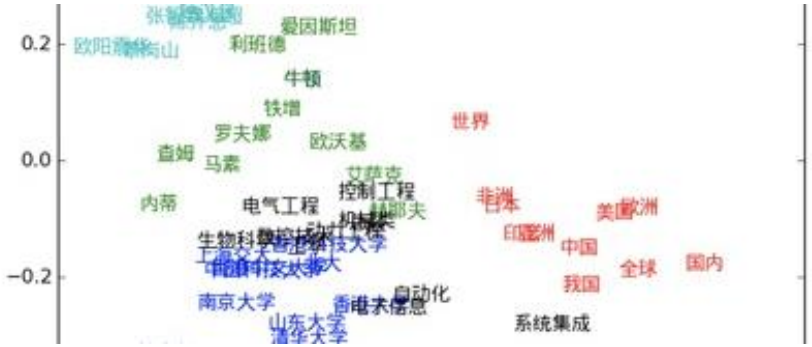
$$\begin{pmatrix} the \\ cat \\ sat \\ on \\ the \\ mat \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

GLoVe
FastText

类比性：



相似性：



算法原理

算法原理

T	使用$d_\chi - privacy$对文本数据进行脱敏
I	输入数个单词组成的字符串
P	1.将输入单词转变为词嵌入向量 2.对词嵌入向量加入满足差分隐私的噪声 3.生成新的扰动词嵌入向量 4.搜索扰动词嵌入向量的最近邻向量，输出最近邻向量对应的单词
O	经过噪声扰动的替换单词字符串
P	单词的空间随着语言词汇表的大小而增长，使得几乎不可能由替换单词近似捕获初始单词的语义
C	1.输入和输出都是文本 2.该机制要能够处理随时间增长的数据集
D	需要证明满足差分隐私变体$d_\chi - privacy$
L	WSDM 2020

算法原理

- 基于文本扰动的$d_\chi - privacy$ 定义
  - 存在随机化机制$M: X \to Y$，对于输出单词 $y \in Y$，任意的输入单词$x, x' \in X$的输出分布满足以下界限：

$$\frac{\Pr(M(x) = y)}{\Pr(M(x') = y)} \leq e^{\varepsilon \cdot d(x,x')}$$

该公式满足了：一对输入单词的词向量越相近，输出的单词就越不容易区分。

  - 其中$d(x, x')$为两个单词的词嵌入向量之间的欧几里得距离（n为词嵌入向量维数）：

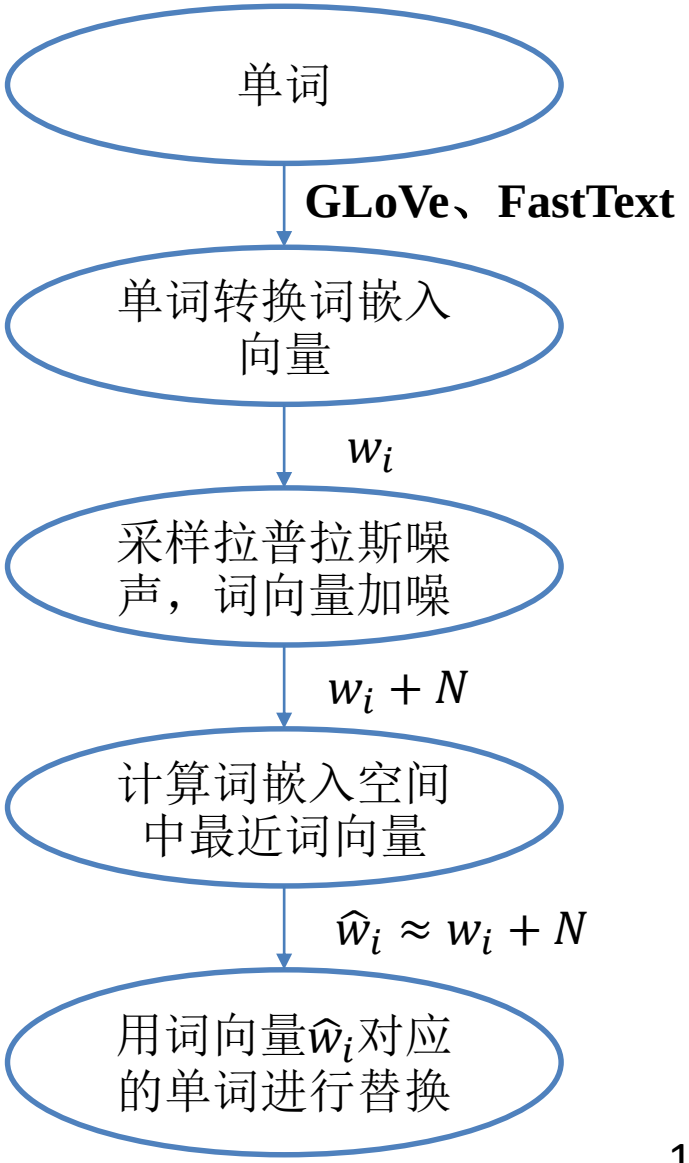
$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x_i')^2}$$

算法原理

- $d_\chi - privacy$ 的随机化机制$M: X \to Y$流程
  - 噪声$N = lv$，其中$v = [v_1, \ldots, v_n]$是从多元正态分布中采样然后归一化得到，l是从伽马分布中采样得到的标量。

- $d_\chi - privacy$ 隐私指标
  - $N_w = \Pr(M(x) = x)$，输入词在机制M中没有被替换的概率。
  - $S_w = |\{y \in W: \Pr(M(x) = y)\} \geq \eta|$，机制$M$输出概率大于$\eta$的不同单词的数量。

N_w越小隐私保护程度越高，S_w相反



算法原理

- 消融实验
  - 实验设置
    - 通过计算输入词w在运行100次机制M后没有被其他单词替换的次数来近似N_w；通过计算输入词w在运行100次机制M后替换的不同单词数量来近似S_w；
  - 结果
    - 随着差分隐私预算ε的增大，隐私保护程度降低，同时N_w上升， S_w下降。

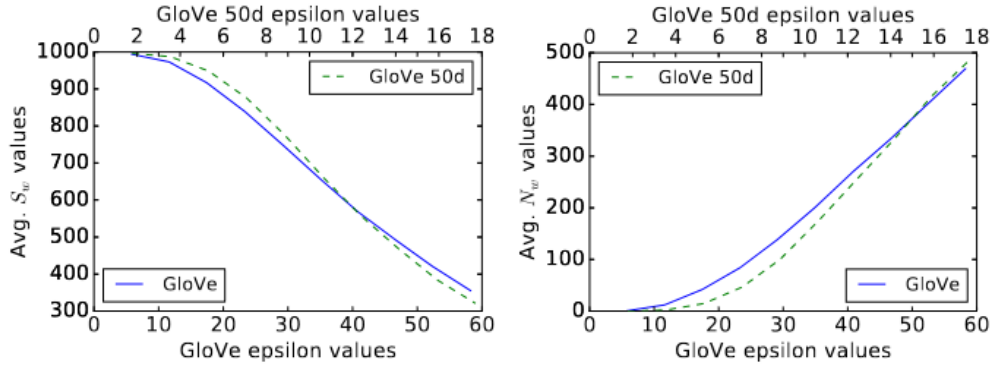


Figure 3: Average S_w and N_w statistics: GloVe 50d and 300d

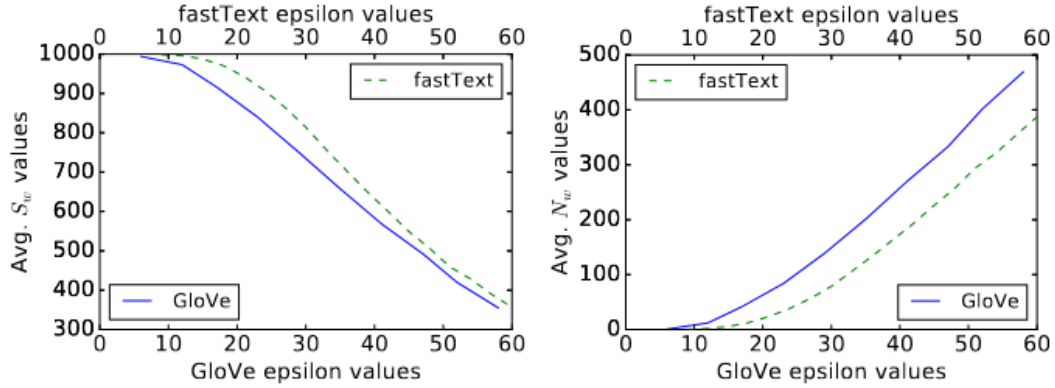


Figure 4: Average S_w and N_w statistics: GloVe and fastText

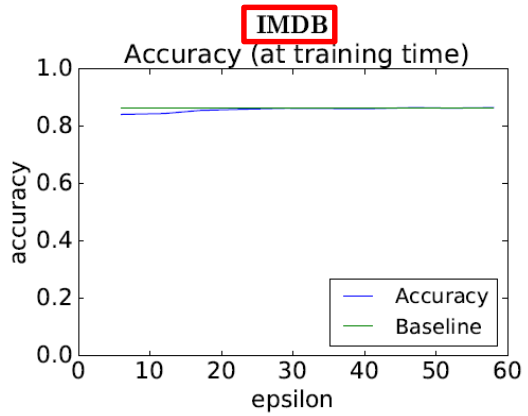
算法原理

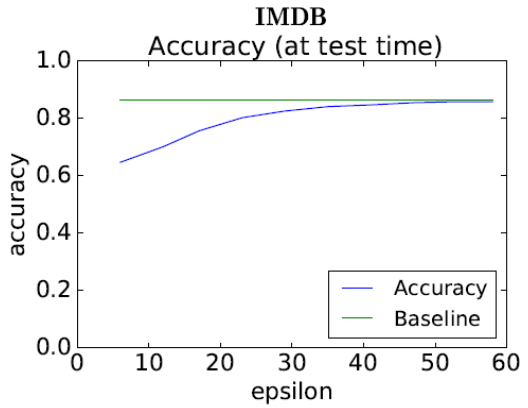
- 数据集
  - IMDb电影评论（IMDb）
    - 判断正面或负面评论，使用分类精度。
  - 安然邮件（Enron）
    - 判断邮件类型属于哪几种类别，使用分类精度。
  - 保险问答（InsuranceQA）
    - 判断文档相似度，使用平均精度（MAP）和平均倒数秩（MRR）。

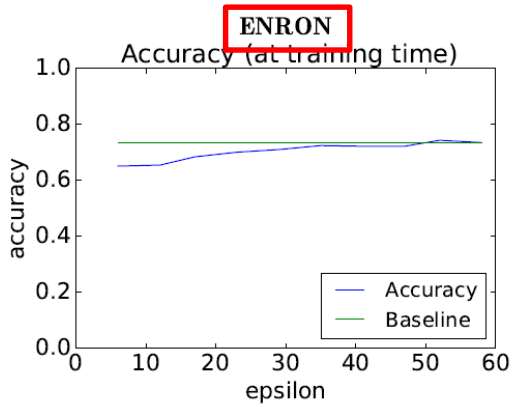
Dataset	IMDb	Enron	InsuranceQA
Task type	binary	multi-class	QA
Training set size	25, 000	8, 517	12, 887
Test set size	25, 000	850	1, 800
Total word count	5, 958, 157	307, 639	92, 095
Vocabulary size	79, 428	15, 570	2, 745
Sentence length	$\mu = 42.27$ $\sigma = 34.38$	$\mu = 30.68$ $\sigma = 31.54$	$\mu = 7.15$ $\sigma = 2.06$

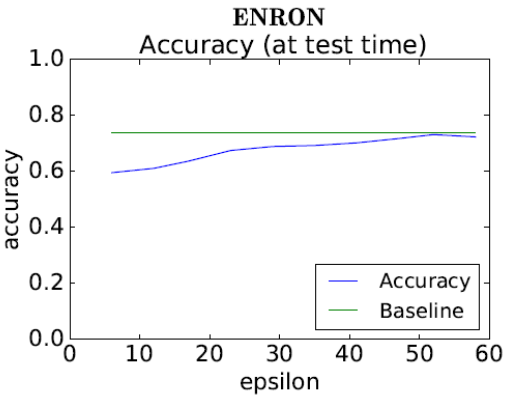
算法原理

- 效用实验
  - 训练阶段：用被扰动的数据训练用于分类的双向LSTM模型，使用该分类模型在普通数据集上进行测试（测试阶段相反）。

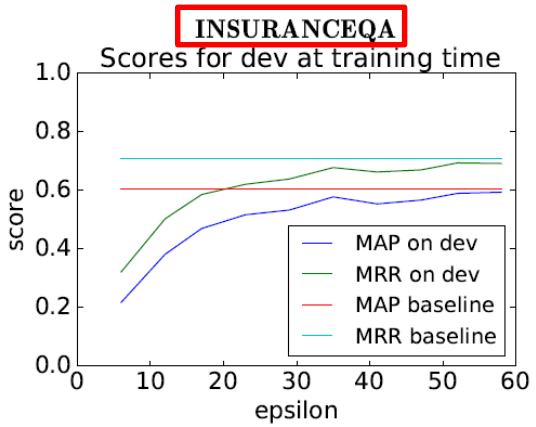


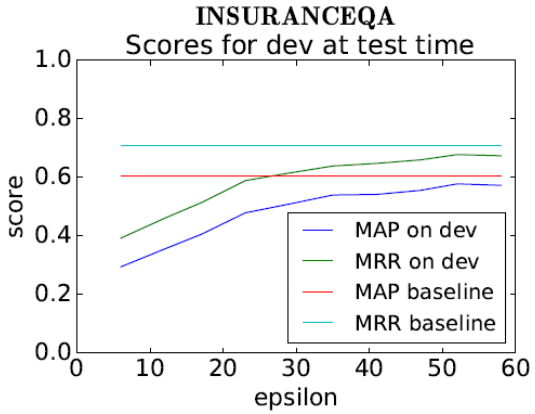






越相似，MAP和MRR越高





算法原理

- 隐私实验
  - 数据集
    - 搜索日志（Search logs）
  - 基线方法
    - Versatile：使用“语义”和“统计”查询扰动技术。
    - Incognito：使用噪音干扰。
  - 实验设置
    - 使用基线方法和隐私化机制M生成扰动查询数据，通过隐私审计系统识别。

Model	Precision	Recall	Accuracy	AUC
Original queries	1.0	1.0	1.0	1.0
Versatile (semantic)	1.0	1.0	1.0	1.0
Versatile (statistical)	1.0	1.0	1.0	1.0
Incognito (without noise)	1.0	1.0	1.0	1.0
Incognito (with noise)	1.0	1.0	1.0	1.0
d_χ-privacy (at $\varepsilon = 23$)	**0.0**	**0.0**	**0.5**	**0.36**

越低隐私保护效果越好

	ε for GloVe 300d								
Metric	6	12	17	23	29	35	41	47	52
Precision	0.00	0.00	0.00	0.00	0.67	0.90	0.93	1.00	1.00
Recall	0.00	0.00	0.00	0.00	0.02	0.09	0.14	0.30	0.50
Accuracy	0.50	0.50	0.50	0.50	0.51	0.55	0.57	0.65	0.75
AUC	0.06	0.04	0.11	0.36	0.61	0.85	0.88	0.93	0.98

算法原理

- 实验结论
  - 随着隐私预算ε的增大，被扰动数据训练的模型效用会越来越贴近基线模型，同时隐私保护程度会下降，体现了隐私-效用的权衡。
- 缺点
  - 有可能会输出原单词；
  - 所加噪声维度过高，会导致替换单词与输入单词语义相似度不高。

		w = encryption		w = hockey		w = spacecraft	
ε	Avg. N_w	GloVe	fastText	GloVe	fastText	GloVe	fastText
⟵ increasing ε, better semantics	50	freebsd multibody 56-bit public-key	ncurses vpns tcp isdn	stadiumarena futsal broomball baseballer	dampener popel decathletes newsweek	telemeter deorbit airbender aerojet	geospace powerup skylab unmanned
	100	ciphertexts truecrypt demodulator rootkit	plaintext diffie-hellman multiplexers cryptography	interleague usrowing football lacrosse	basketball lacrosse curlers usphl	laser apollo-soyuz agena phaser	voyager cassini-huygens adrastea intercosmos
	200	harbormaster unencrypted cryptographically authentication	cryptographic ssl/tls authentication cryptography	players ohl goaltender defenceman	goaltender ephl speedskating eishockey	launch shuttlecraft spaceborne interplanetary	orbited tatooine flyby spaceborne
	300	decryption encrypt encrypted encryption	encrypt unencrypted encryptions encrypted	nhl hockeydb hockeyroos hockey	hockeygoalies hockeyroos hockeyettan hockey	spaceplane spacewalk spaceflights spacecraft	spaceship spaceflights satellites spacecraft

算法原理

T	对文本进行净化
I	文本文档
P	1.输入包含L个单词的文档 2.把一个单词映射为词嵌入向量 3.以概率公式获取一个词嵌入向量替换当前单词 4.循环步骤2和步骤3直到所有单词被替换
O	生成经过净化处理的文本文档（脱敏文档）

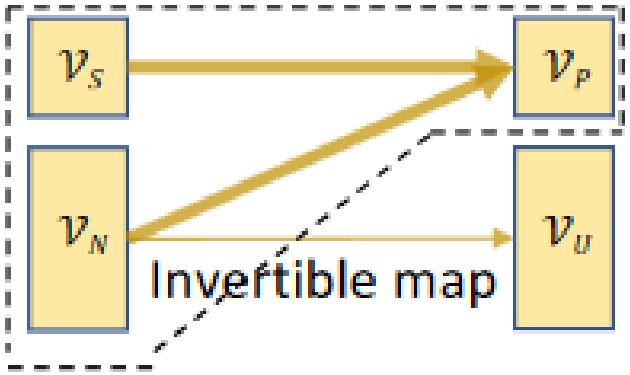
P	NLP工作只关注于特征或文本表示，使得透明度值得怀疑
C	将公共数据当作训练数据
D	可证明和可量化的隐私保障下保护文本数据的效用
L	ACL 2021

算法原理

- 度量局部差分隐私（MLDP）
  - 给定差分隐私参数$\epsilon > 0$，和一个距离度量$d: \mathcal{V} \times \mathcal{V} \to R_{\geq 0}$，$\mathcal{V}$是包含$|\mathcal{V}|$个单词的词汇表，$M$满足$MLDP$或$\epsilon \cdot d(x, x') - LDP$如果，对于任何$x', x, y \in \mathcal{V}$，

$$Pr[M(x) = y] \leq e^{\epsilon \cdot d(x, x')} \cdot Pr[M(x') = y]$$

- 效用局部差分隐私（ULDP）
  - $\mathcal{V}_S \subseteq \mathcal{V}$是所有用户通用的敏感单词集，$\mathcal{V}_N = \mathcal{V} \backslash \mathcal{V}_S$是剩余单词集。将输出空间$\mathcal{V}$分割为受保护部分$\mathcal{V}_P \subseteq \mathcal{V}$，和不受保护部分$\mathcal{V}_U \subseteq \mathcal{V} \backslash \mathcal{V}_P$。



本篇论文中定义低频的单词为敏感单词

算法原理

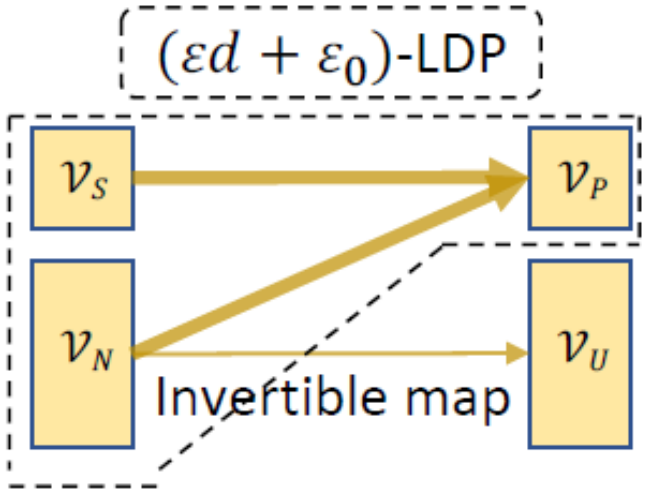
- 效用度量差分隐私（UMLDP）
  - 给定差分隐私参数$\epsilon > 0$，和一个距离度量$d: \mathcal{V} \times \mathcal{V} \to R_{\geq 0}$，$M$满足$UMLDP$，如果

（1）对于任何$x', x, y \in \mathcal{V}$，都有

$$Pr[M(x) = y] \leq e^{\epsilon \cdot d(x,x') + \epsilon_0} \cdot Pr[M(x') = y]$$

（2）对于任何$y \in \mathcal{V}_U$，例如未受保护的组$\mathcal{V}_U$其中$\mathcal{V}_U \cap \mathcal{V}_P = \emptyset$，存在一个$x \in \mathcal{V}_N$使

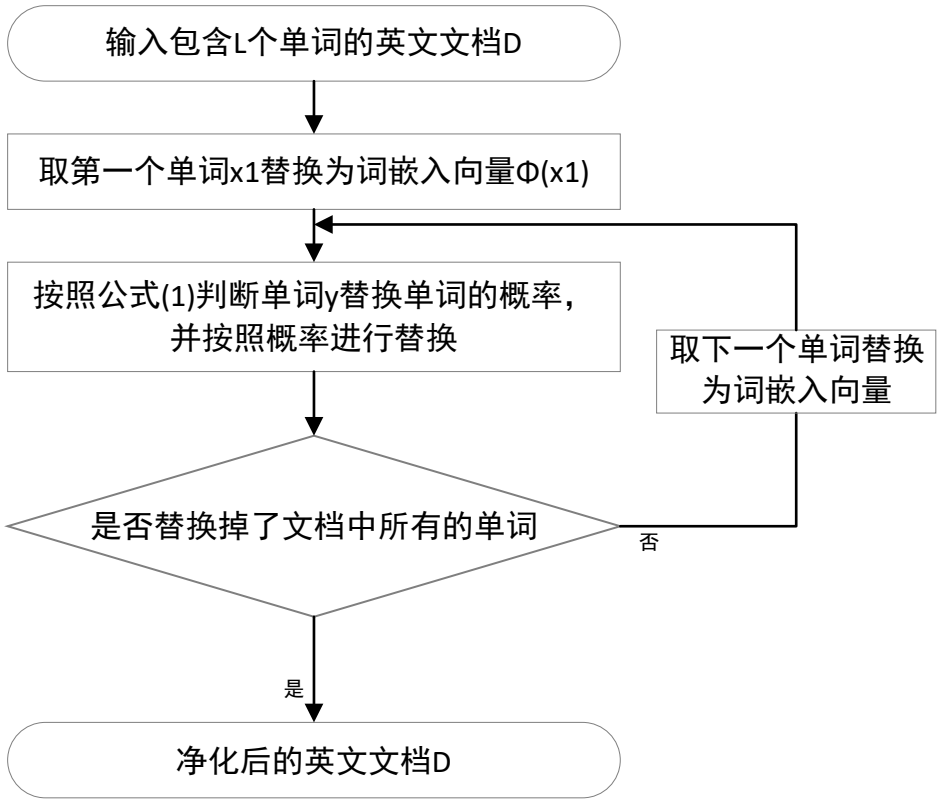
$$\Pr[M(x) = y] > 0, \Pr[M(x') = y] = 0 \forall x' \in \mathcal{V} \backslash \{x\}$$



算法原理

- **算法1：SANTEXT**

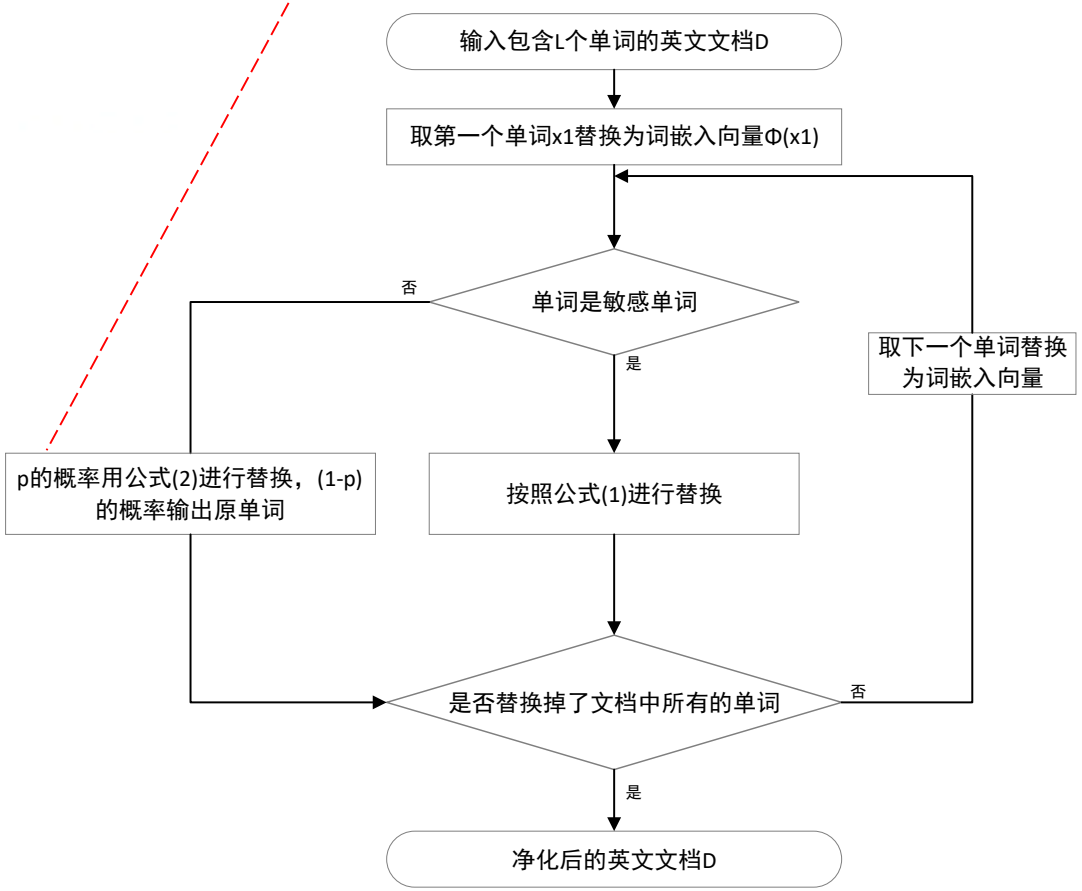
$$Pr[M(x) = y] = C_x \cdot e^{-\frac{1}{2}\epsilon\, d_{euc}(\phi(x),\phi(y))} \qquad (1)$$



算法原理

- **算法2： SANTEXT+**

$$Pr[M(x) = y] = p \cdot C_x \cdot e^{-\frac{1}{2} d_{euc}(\phi(x),\phi(y))} \quad (2)$$



算法原理

- 数据集
  - 情绪分类（SST-2）
    - 判断正面评论和负面评论，使用分类精度。
  - 医学语义文本相似性（MEDST）
    - 判断患者病例相似性，使用分类精度。
  - 自然语言推理（QNLI）
    - 判断给定的文档是否包含问题的答案，使用分类精度。

Datasets	GloVe embeddings		BERT embeddings	
	$\mathcal{V}$	$\mathcal{V}_S$	$\mathcal{V}$	$\mathcal{V}_S$
SST-2	14,730	13,258	30,522	27,469
MedSTS	3,320	2,989		
QNLI	88,159	79,343		

总单词集合:敏感单词集合=10:9

算法原理

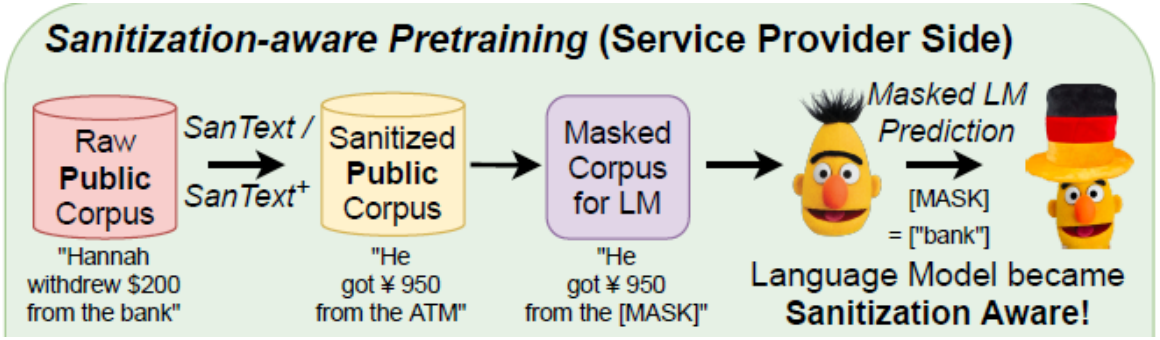
- 对比实验
  - 使用机制M对SST-2等数据集进行扰动，用输出的扰动数据集训练分类模型，计算该模型在原始数据集上的准确率。

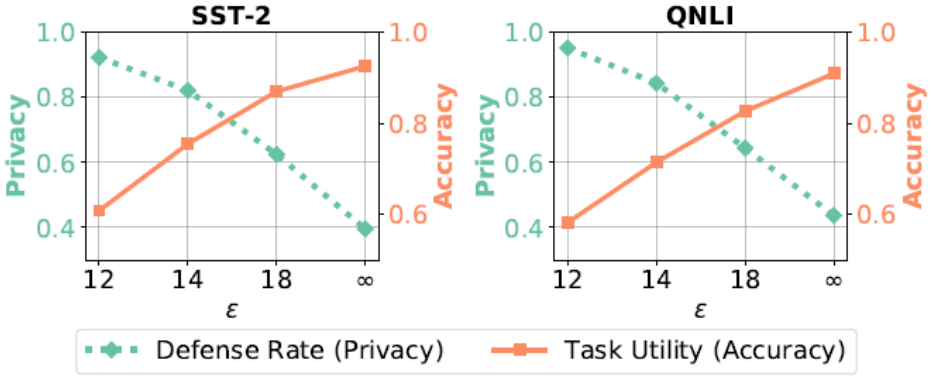
Mechanisms	SST-2			MedSTS			QNLI		
	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
Random	0.4986	0.4986	0.4986	0.0196	0.0196	0.0196	0.5152	0.5152	0.5152
Feyisetan et al. (2020)	0.5099	0.5143	0.5345	0.0201	0.0361	0.0452	0.5162	0.5256	0.5333
SANTEXT	0.5101	0.5838	0.8374	0.0351	0.5392	0.8159	0.5372	0.5598	0.8116
SANTEXT$^+$	**0.7796**	**0.7943**	**0.8516**	**0.4965**	**0.7082**	**0.8162**	**0.7699**	**0.7760**	**0.8131**
Unsanitized	0.9251			0.8527			0.9090		

$d_\chi - privacy$

算法原理

- 消融实验
  - 脱敏单词推理攻击实验
    - 对经过净化的文本，依次用特殊的单词[MASK]替换每个单词，并将文本输入到BERT模型中，得到[MASK]所在位置的单词的预测，然后将预测的单词与原始文本中的单词进行比较。

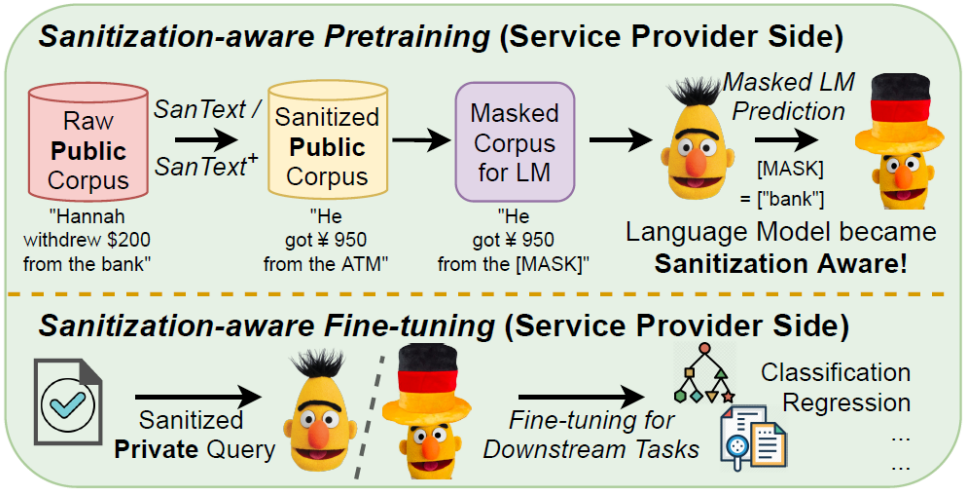




隐私性：不匹配单词占总单词的比例
准确率：用作数据集，训练出来的模型的准确率

算法原理

- 消融实验
  - 预训练的影响
    - 净化感知模型在*bert-base-uncased*模型的基础上用维基百科语料库的扰动版本进行预训练，然后使用掩码语言模型（Masked Language Model）进行预训练。
    - 比较了SST-2和QNLI在不同隐私级别下，基于原始公开的*bert-base-uncased*模型和净化感知模型的准确性。



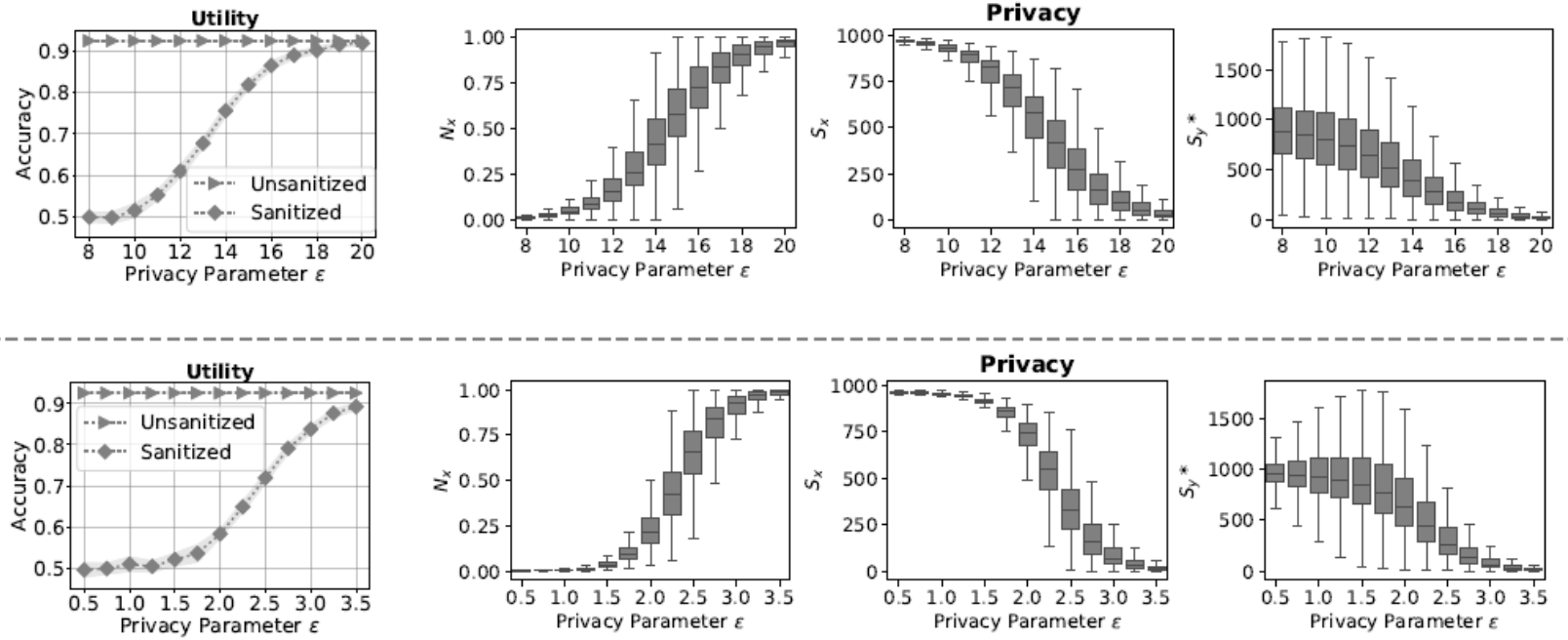
Datasets	ϵ	Utility		Δ_{privacy}
		Original	+Pretrain	
SST-2	12	0.6084	**0.6208**	0.0089
	14	0.7548	**0.7731**	0.0101
	16	0.8698	**0.8830**	-0.0046
QNLI	12	0.5822	**0.6037**	0.0076
	14	0.7143	**0.7309**	-0.0047
	16	0.8265	**0.8369**	-0.0039

算法原理

- 消融实验
  - 使用了三个指标来量化隐私：N_x、S_x、S_y^*

一组输入集合x可能以相似的概率输出y，定义S_y^*为该集合的大小

  - 随着差分隐私预算ε的增大，隐私保护程度降低，同时N_x上升， S_x和S_y^*下降。



算法原理

- 实验结论
  - 通过概率替换单词的方式避免了添加高维噪声，相同隐私预算下效用高于$d_\chi - privacy$，且用净化语料库预训练的Bert模型效用更高，隐私保护程度也不一定会下降。
- 缺点
  - 有可能输出原单词；

Dataset: SST-2		
Mechanisms	ϵ	**Original Text:** it 's a charming and often affecting journey .
SANTEXT	1	heated collide. charming activity cause challenges beneath tends
	2	worse beg, charming things working noticed journey basically
	3	all 's. charming and often already journey demonstrating
SANTEXT$^+$	1	it unclear a charming and often hounds journey
	2	it exaggeration a charming feelings often lags journey .
	3	it 's a tiniest picked often affecting journey .

应用总结

- 大数据脱敏系统
  - 敏感信息脱敏
  - 个人隐私保护
  - 防止实体关系抽取

参考文献

- [1] Feyisetan O, Balle B, Drake T, et al. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 178-186.

- [2] Yue X, Du M, Wang T, et al. Differential Privacy for Text Analytics via Natural Text Sanitization[J]. arXiv preprint arXiv:2106.01221, 2021.

知人者智，自知者明。

胜人者有力，自胜者强。

知足者富，强行者有志。

不失其所者久，死而不亡者，寿。

谢谢！