A Linguistics-based Approach for Achieving Sentence-level Differential Privacy

Chaeeun (Joy) Lee

22.04.2024, Bachelor Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

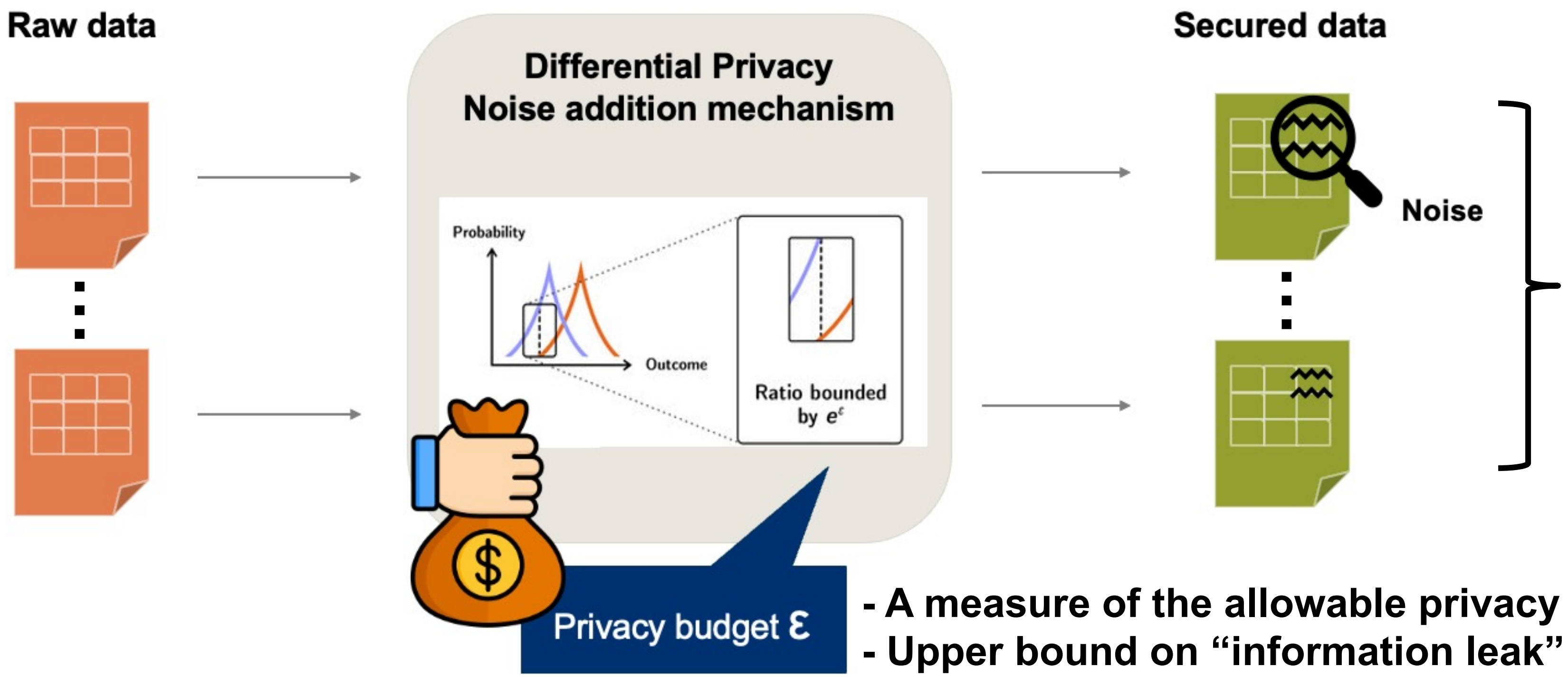
1. Motivation & Research Questions
2. Methodology
3. Result & Key Findings
4. Conclusion

Motivation

RECAP

What is DP?

Privacy-enhancing technique provides strong privacy guarantees by introducing controlled noise to individual data points Dwork et al. (2006, "Differential Privacy")



obscuring individual contributions to prevent identification

- A measure of the allowable privacy loss
- Upper bound on “information leak”

Image : Franzen, Daniel & Nuñez von Voigt, Saskia & Sörries, Peter & Tschorsch, Florian & Müller-Birn, Claudia. (2022). "Am I Private and If So, how Many?" -- Using Risk Communication Formats for Making Differential Privacy Understandable.

© sebis

Motivation

RECAP

Conventional (word-level) approach

Applied to each individual word in the sentence equally

She enjoys reading novels in her cozy, quiet room.

$\varepsilon = 1.0$

[She] [enjoys] [reading] [novels] [in] [her] [cozy] [quiet] [room]

↓ 0.1 ↓ 0.1 … ↓ 0.1 ↓ 0.1

[He] [delights] [devouring] [books] [within] [his] [snug] [tranquil] [space]

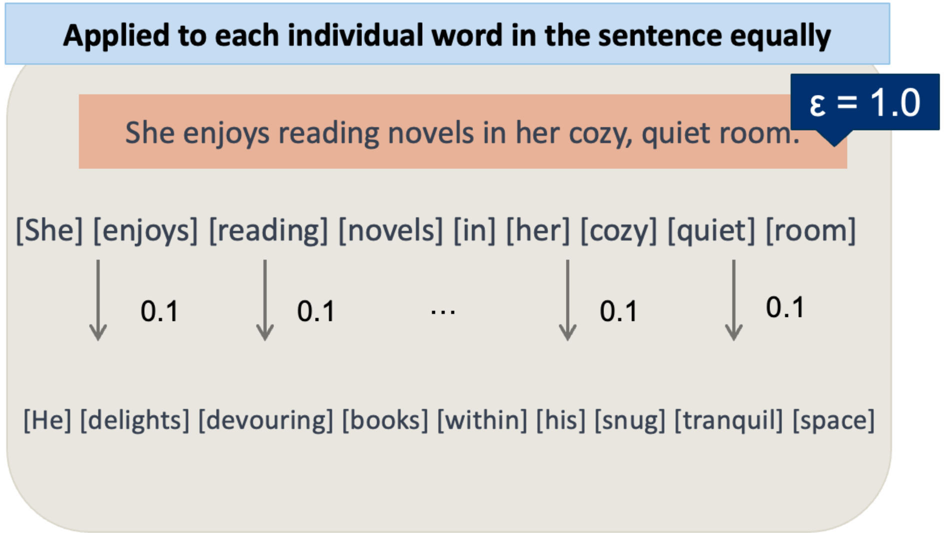
⚠ naive distribution of the budget

© sebis

Motivation

RECAP

Conventional (word-level) approach



⚠ naive distribution of the budget

What is the reasonable way to distribute the limited privacy budget to achieve sentence-level DP?

Motivation

RECAP

New approach - Sentence-Level

What is the reasonable way to distribute the limited privacy budget to achieve sentence-level DP?

Conventional (word-level) approach

Applied to each individual word in the sentence equally

ε = 1.0

She enjoys reading novels in her cozy, quiet room.

[She] [enjoys] [reading] [novels] [in] [her] [cozy] [quiet] [room]

0.1 0.1 … 0.1 0.1

[He] [delights] [devouring] [books] [within] [his] [snug] [tranquil] [space]

⚠ naive distribution of the budget

Sentence-Level Privacy with linguistics-based analysis

ε = 1.0

She enjoys reading novels in her cozy, quiet room.

[She] [enjoys] [reading] [novels] [in] [her] [cozy] [quiet] [room]

0.05 0.3 … 0.2 0.15

???

Informativeness as the criteria:

A word containing more information in the text is more likely to be significant for identification and that it needs to be protected.

Research Questions

RECAP

RQ1

How can DP be effectively applied at the sentence level within Natural Language Processing, considering the intelligent distribution of privacy budgets for individual words within a sentence?

RQ2

How can the theoretical concept of sentence-level privacy with informativeness analysis be translated into an implementable framework?

RQ3

How well does the suggested differential privacy framework protect private data while preserving the utility of the text data?

© sebis

Methodology - Overview

RQ1

Prototype Implementation

Design and develop an ε Distributor, incorporating linguistic methods to allocate the privacy budget to each word in the text

- Input: text, total budget
- Consist of 5 informativeness scoring methods
- Calculate the final budget by combining the scores
- Output: distributed budget of each word

```
[('After', 0.010595947716120033),
 ('graduating', 0.033979150676310604),
 ('from', 0.007747180331524092),
 ('Stanford', 0.054382610929507275),
 ('University', 0.06041720513544317),
 ('John', 0.08443002084143324),
 ('Smith', 0.08509501518205213),
```

RQ2

Dataset Construction

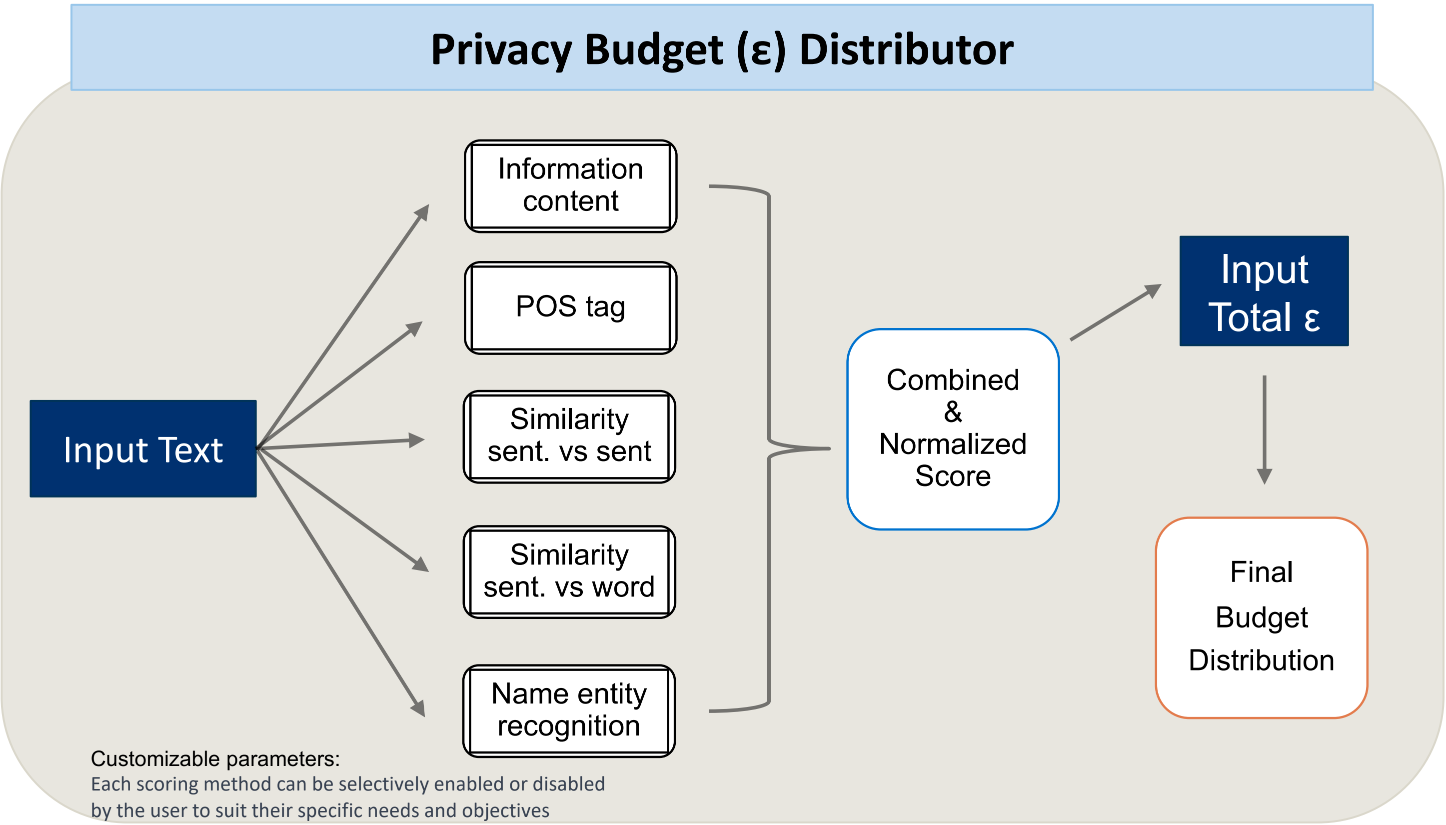
RQ3

Experiment & Evaluation

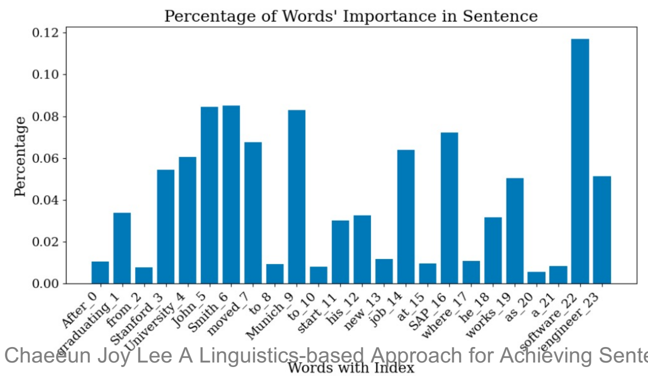
*GLUE: General Language Understanding Evaluation

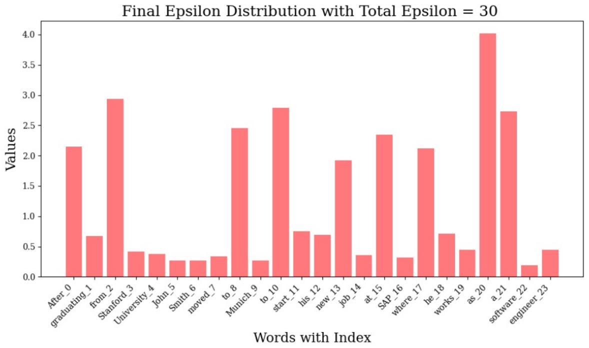
Methodology – Prototype Implementation

RECAP



Token	IC	POS	NER	Sentence Sim.	Word Sim.	Final Score	ϵ
After	1.0	0.1	0	0.0176	0.0510	0.0106	2.1480
graduating	185.17	8	0	0.0195	0.2069	0.0340	0.6698
from	1.0	0.1	0	0.0115	0.0431	0.0077	2.9379
Stanford	44.9	15	1	0.0298	0.3098	0.0544	0.4185
University	7410.33	15	1	0.0180	0.2483	0.0604	0.3767
John	17607.56	15	1	0.0492	0.1420	0.0844	0.2696
Smith	1740.58	15	1	0.1129	0.2719	0.0851	0.2675
moved	16475.69	8	0	0.0317	0.1485	0.0675	0.3373
to	1.0	0.1	0	0.0135	0.0534	0.0093	2.4536
Munich	129.2	15	1	0.1239	0.2350	0.0828	0.2749
start	5149.93	8	0	0.0111	0.0703	0.0303	0.7510
his	1.0	14	0	0.0167	0.1135	0.0326	0.6971
new	1.0	3.7	0	0.0138	0.0276	0.0119	1.9199
job	14954.66	15	0	0.0132	0.1162	0.0638	0.3568
at	10.4	0.1	0	0.0143	0.0551	0.0097	2.3423
SAP	300.48	15	1	0.0679	0.3545	0.0723	0.3148
where	1.0	0.1	0	0.0140	0.0693	0.0107	2.1205
he	135.9	14	0	0.0153	0.1048	0.0317	0.7177
works	17173.41	8	0	0.0100	0.0169	0.0505	0.4505
as	53.84	0.1	0	0.0131	0.0084	0.0057	4.0204
a	48.2	0.1	0	0.0116	0.0491	0.0083	2.7288
software	37852.6	15	0	0.0250	0.1338	0.1169	0.1948
engineer	1549.1	15	0	0.0233	0.2604	0.0512	0.4446





Example result of ε distributor prototype using an example sentence "After graduating from Stanford University, John Smith moved to Munich to start his new job at SAP, where he works as a software engineer" and the 30 total epsilon.

Methodology - Overview

Prototype Implementation (RQ1)

Design and develop an ε Distributor, incorporating linguistic methods to allocate the privacy budget to each word in the text

- Input: text, total budget
- Consist of 5 informativeness scoring methods
- Calculate the final budget by combining the scores
- Output: distributed budget of each word


```
[('After', 0.010595947716120033),
 ('graduating', 0.033979150676310604),
 ('from', 0.007747180331524092),
 ('Stanford', 0.054382610929507275),
 ('University', 0.060417205135443317),
 ('John', 0.08443002084143324),
 ('Smith', 0.08509501518205213),
```


Dataset Construction (RQ2)

Choose the dataset and perturb the text data using 2 DP mechanisms w/ or w/o the prototype

- 2 DP mechanisms
  - 1-Diffractor: based on word-level Metric Local Differential Privacy (MLDP) mechanisms
  - DP-MLM: leverages masked token prediction in BERT-based models
- Privacy: Trustpilot (gender), Yelp (user id)
- Utility: GLUE* Benchmark - CoLA, SST-2, MRPC, RTE, STSB
- Perturb each text in the dataset with or without the distributor using mechanisms

Experiment & Evaluation (RQ3)

*GLUE: General Language Understanding Evaluation

© sebis

Methodology – Dataset Construction Pipeline

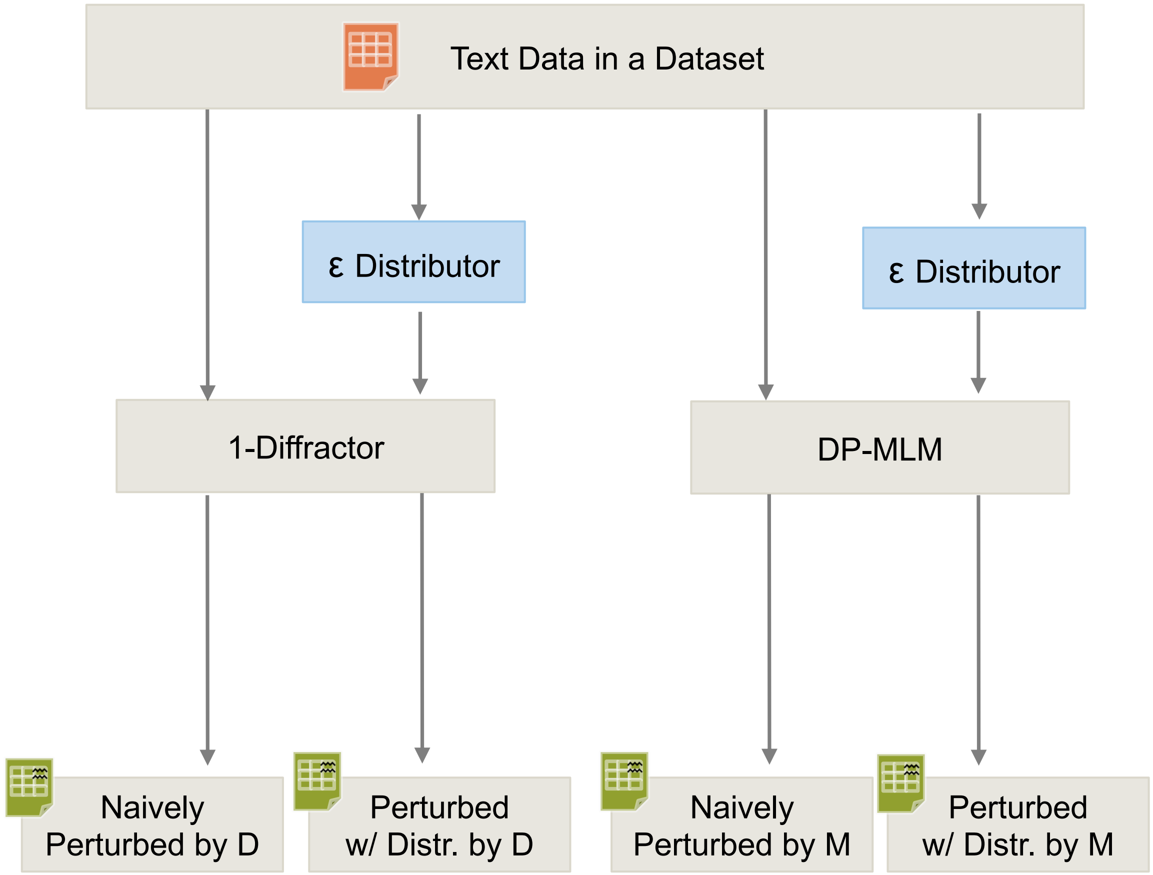


Table of datasets and the standard ε value used in this thesis.

Type	Dataset	Size	Metric	Avg. word count	Total ϵ (1-Diffractor)	Total ϵ (DP-MLM)
Privacy	Trustpilot	36621	Accuracy	45	45	4500
	Yelp	17336	Accuracy	182	182	18200
Utility	CoLA	8551/1043	Accuracy	8	8	800
	SST-2	30000/872	Accuracy	9	9	900
	MRPC	3668/408	Accuracy & F1 Score	22	22	2200
	RTE	2490/277	Accuracy	43	43	4300
	STSB	5749/1500	Pearson-Spearman correlation	10	10	1000

Example of perturbed dataset (CoLA dataset)

	sentence	label	naive_dp_sentence_M	distributed_dp_sentence_M
0	Our friends won't buy this analysis, let alone...	1	Your friends wo not love this analysis, left i...	Your pals wo 't buy this analysis, let alone t...
1	One more pseudo generalization and I'm giving up.	1	One more pseudo general and O're failing up	Used more fake spectrum and He mean catching
2	One more pseudo generalization or I'm giving up.	1	No more pseudo general or You am giving up	So more pseudo roundup or Me're telling
3	The more we study verbs, the crazier they get.	1	Athe more we manipulate verbs, the tighter the...	So more we understand pronouns, the darker they
4	Day by day the facts are getting murkier.	1	Game by everyday the probabilities are dying w...	Hopefully by week the stats are breaking weaker

© sebis

Methodology - Overview

RQ1 — Prototype Implementation

Design and develop an ε Distributor, incorporating linguistic methods to allocate the privacy budget to each word in the text

- Input: text, total budget
- Consist of 5 informativeness scoring methods
- Calculate the final budget by combining the scores
- Output: distributed budget of each word

```
[('After', 0.010595947716120033),
 ('graduating', 0.033979150676310604),
 ('from', 0.007747180331524092),
 ('Stanford', 0.054382610929507275),
 ('University', 0.060417205135443l7),
 ('John', 0.08443002084143324),
 ('Smith', 0.08509501518205213),
```

RQ2 — Dataset Construction

Choose the dataset and perturb the text data using 2 DP mechanisms w/ or w/o the prototype

- 2 DP mechanisms
  - 1-Diffractor: based on word-level Metric Local Differential Privacy (MLDP) mechanisms
  - DP-MLM: leverages masked token prediction in BERT-based models
- Privacy: Trustpilot (gender), Yelp (user id)
- Utility: GLUE* Benchmark - CoLA, SST-2, MRPC, RTE, STSB
- Perturb each text in the dataset with or without the distributor using mechanisms

RQ3 — Experiment & Evaluation

Analyze privacy & utility evaluation results

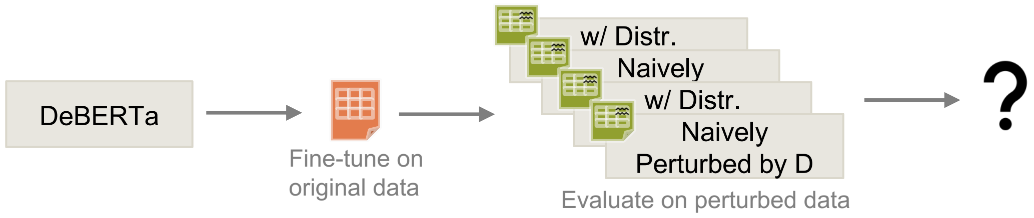
- **Finetune** a pre-trained model (DeBERTa) and **evaluate**:
- Compare the result of each evaluation metric (Accuracy, F1 score …)
- **Privacy: How well does the model predict certain characteristics of individual data?**
  Accuracy ↓ => Privacy ↑
- **Utility: How much does the rewritten dataset affect the NLU performance of the model?**
  Metric ↑ = Utility ↑

*GLUE: General Language Understanding Evaluation

Methodology – Evaluation process

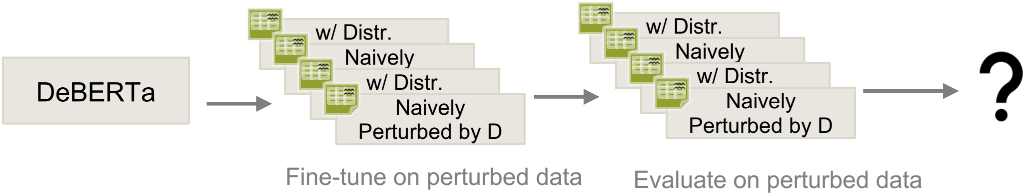
Main experiment - Privacy

- Fine-tune DeBERTa-v3-base on **original texts** in the dataset
- Evaluate the model with perturbed texts and compare the result
- Label : Trustpilot – gender(2) / Yelp – user id (10)
- Metric : Accuracy



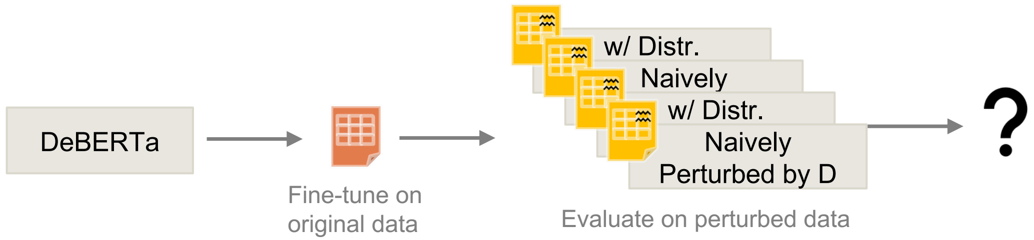
Main experiment - Utility

- Fine-tune DeBERTa-v3-base on each **perturbed training dataset**
- Evaluate the model with the perturbed evaluation dataset
- Label : 2 except STSB (continues value)
- Metric : Accuracy, F1-score, Pearson- Spearman correlation



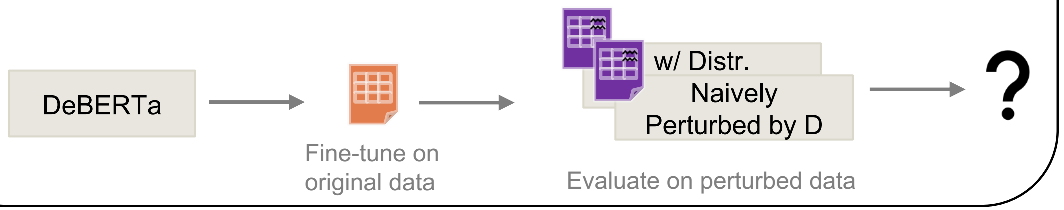
Sub-experiment - Stop-word Filtering

- Privacy evaluation comparison on datasets perturbed **without the stop-words filtering option** of the DP mechanisms
- Trustpilot with stop-word filter disabled 1-Diffractor, DP-MLM



Sub-experiment - Word-level Privacy Budget application

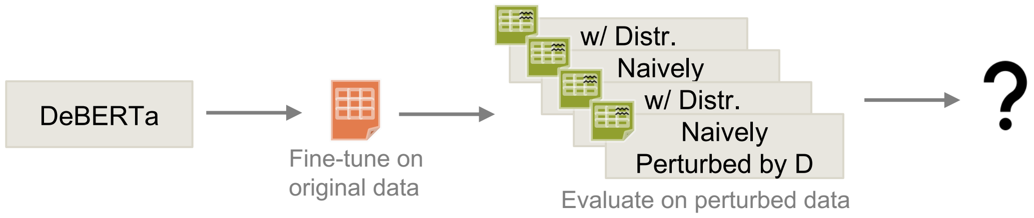
- Privacy evaluation comparison on datasets perturbed **with individual privacy budgets** (each data point gets a different privacy budget based on the size of its text)
- To show the impact of the Distributor in word-level budget setting
- Trustpilot & Yelp with 1-Diffractor



Methodology – Evaluation process 1

Main experiment - Privacy

- Fine-tune DeBERTa-v3-base on **original texts** in the dataset
- Evaluate the model with perturbed texts and compare the result
- Label: Trustpilot – gender(2) / Yelp – user id (10)
- Metric: Accuracy

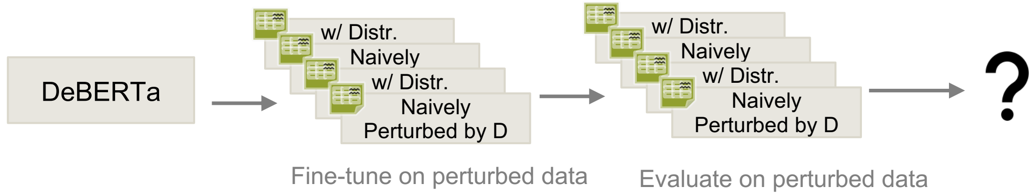


Sample example: Trustpilot dataset perturbed with 1-Diffractor

	text	gender	naive-d	distributed-d
0	Found my favourite pen!!!: I have been using t...	F	Found my favourite pens I have been used this ...	Found my favorites pens I have been using this...
1	poor customer service: Receive part in a box t...	M	attain patrons restricted Receive part in a bo...	impoverished diners servicing Receive their in...
2	av		awesome Best prices EVAR	awesome Best prices EVAR
3	best products for tl		best brands for the price I always order for m...	most product for the bidder I thing ordering f...
4	quick and easy: I used		quick and turbo I taught Rush My Passport in o...	reg and viable I utilizes Rush My Passport in ...

Main experiment - Utility

- Fine-tune DeBERTa-v3-base on each **perturbed training dataset**
- Evaluate the model with the perturbed evaluation dataset
- Label: 2 except STSB (continues value)
- Metric : Accuracy, F1-score, Pearson- Spearman correlation



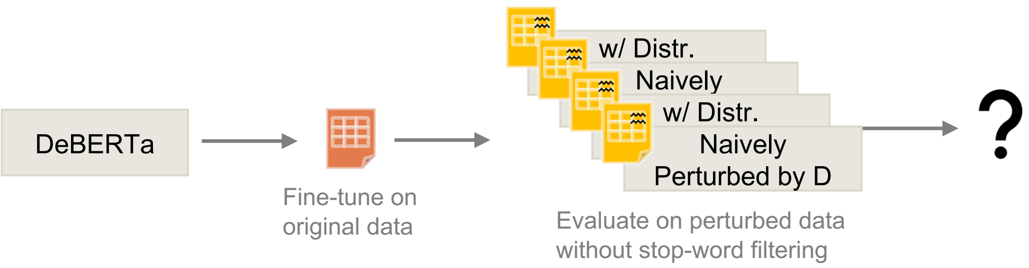
Sample example: MRPC dataset perturbed with DP-MLM

	sentence1	sentence2	label	naive1-m	distributed1-m	naive2-m	distributed2-m
0	He said the foodservice pie business doesn 't ...	" The foodservice pie business does not fit ou...	1	She added the service pie line doesn t satisfy...	Ceo added the snack pie segment doesn t captur...	A service pie company does not join our long e...	Our service pie businesses does not fitting ou...
1	Magnarelli said Racicot hated the Iraqi regime...	His wife said he was " 100 percent		He explained He fled the Iraqi	Cade stated Creep admired the Present torture ...	He wife friday he was 50 percent with Gore Bus...	Former tourist says he was ten completely behi...
2	The dollar was at 116.92 yen against the yen ,...				Japanese dollar was at 465 counter against the...	The dollar was at 1911 yen Jp essentially cons...	Nz dollar was at 1100 he, largely flat on the ...
3	The AFL-CIO is waiting until October to decide...				The Afl is holding until November to pick if i...	* Io outlined Wednesday that it will decided i...	The Nfl tweeted Today that it will see in July...
4	No dates have been set for the civil or the cr...				Battle where have been set for the civil or th...	Neither months have been schedule for the crim...	No noses have been sat for the criminal or sex...

Methodology – Evaluation process 2

Sub-experiment - Stop-word Filtering

- Privacy evaluation comparison on datasets perturbed **without the stop-words filtering option** of the DP mechanisms
- Trustpilot with stop-word filter disabled 1-Diffractor, DP-MLM

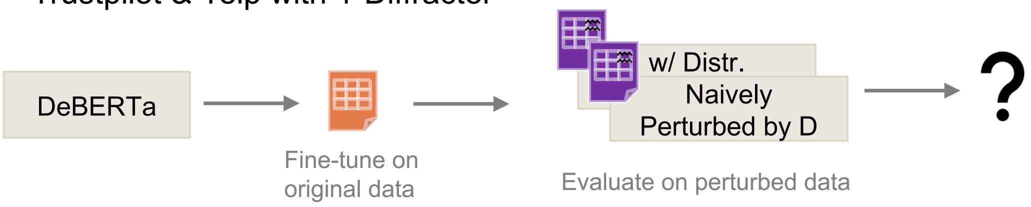


Sample example:

Original	Stefan is studying in Germany
Perturbed w/ stop-word filtering (default)	He is learning in German
Perturbed w/o stop-word filtering	She was looking under Germany
Perturbed w/o stop-word filtering w/ ε distributor	Ryan is succeeding in Berlin

Sub-experiment - Word-level Privacy Budget application

- Privacy evaluation comparison on datasets perturbed **with individual privacy budgets** (each data point gets a different privacy budget based on the size of its text)
- To show the impact of the Distributor in word-level budget setting
- Trustpilot & Yelp with 1-Diffractor



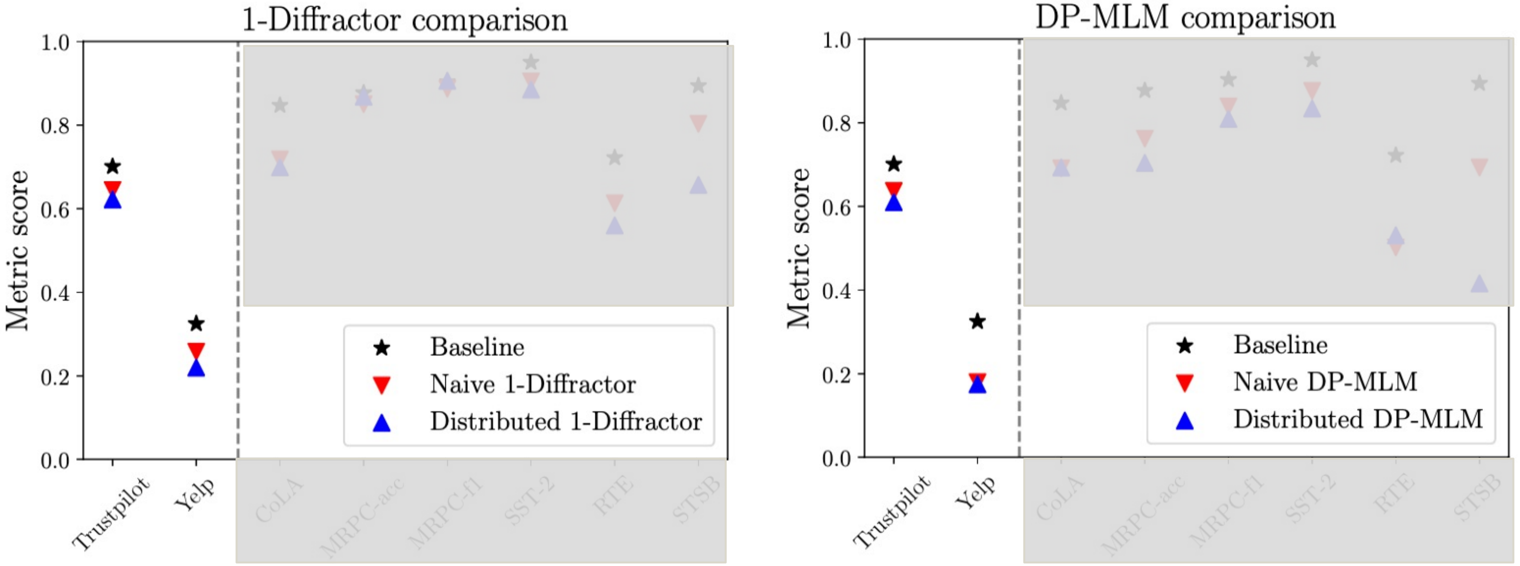
Sample example:

Original	Stefan is studying in Germany (length: 5)	The future belongs to those who believe in the beauty of their dreams (length: 14)
perturbed w/ fixed budget (default)	[budget: 9.5] She is reading in Germany	[budget: 9.5] That future presents to those who faith in the ere of their better
Perturbed w/ individual budget	[budget: 5] He is looking in Berlin	[budget: 14] The future maps to those who see in the majesty of their dreams

© sebis

Result & Key Findings - The consistent improvement of privacy

Main experiment – Privacy



Visualization of the main privacy experiment

- Consistently enhanced privacy preservation (lower accuracy) resulted from both DP mechanisms.
- Enhanced privacy (lower accuracy) in sub-experiments; both stop-word filtering and word-level budget application

Sub-experiment - Stop-word Filtering

Dataset	Baseline	1-Diffractor				DP-MLM			
		budget	naive	ϵ-distr.	diff.	budget	naive	ϵ-distr.	diff.
Trustpilot (Ref.)	0.693	45	0.645	0.622	−0.023	4500	0.637	0.610	−0.027
Trustpilot (Stop)	0.693	45	0.628	0.612	−0.016	4500	0.584	0.581	−0.003
Trustpilot (Stop 1/2)	0.693	22	0.595	0.576	−0.019	2200	0.579	0.562	−0.017

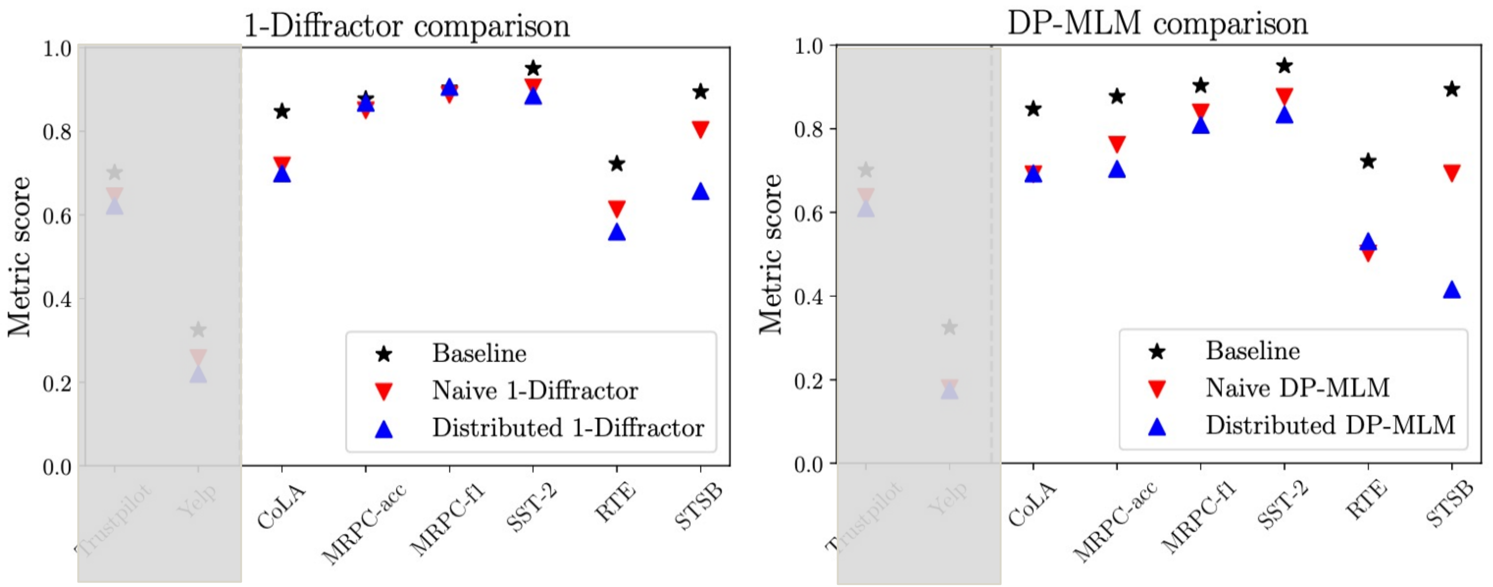
Sub-experiment - Word-level Privacy Budget application

Dataset	Baseline	Individual budget			
		budget	naive	ϵ-distr.	diff.
Trustpilot	0.693	len(text)	0.671	0.618	−0.053
Yelp	0.325	len(text)	0.303	0.195	−0.108

Evaluation results of two sub-experiments

Result & Key Findings - Maintenance and loss of utility

Main experiment – Utility



Visualization of the main utility experiment.

- The utility has been maintained - the similar performance scores observed across the datasets
  (1-Diffractor: MRPC,
  DP-MLM: CoLA, RTE)
- Utility scores decrement in certain datasets and with specific differential privacy mechanisms
  (1-Diffractor: CoLA, SST-2, RTE
  DP-MLM: MRPC, SST-2)
- Noticeable utility loss
  (STSB)

Result & Key Findings - Further insights on budget choice and stop-word filtering

Sub experiment - Stop-word Filtering

Dataset	Baseline	1-Diffractor				DP-MLM			
		budget	naive	ϵ-distr.	diff.	budget	naive	ϵ-distr.	diff.
Trustpilot (Ref.)	0.693	45	0.645	0.622	−0.023	4500	0.637	0.610	−0.027
				V				V	
Trustpilot (Stop)	0.693	45	0.628	0.612	−0.016	4500	0.584	0.581	−0.003
					Λ				Λ
Trustpilot (Stop 1/2)	0.693	22	0.595	0.576	−0.019	2200	0.579	0.562	−0.017

Evaluation result of the sub-experiment.
Trustpilot (Stop) is perturbed with the stop-word filtering option disabled.
Trustpilot (Stop 1/2) is perturbed with the stop-word filtering option disabled, using half of the standard privacy budget.

- Improved (lower accuracy) privacy when the stop-word filtering is disabled
- The more limited the budget, the more difference there was in improving privacy (larger difference)

Sub experiment - Word-level vs. Sentence-level Privacy Budget

Dataset	Baseline	Individual budget				Fixed budget			
		budget	naive	ϵ-distr.	diff.	budget	naive	ϵ-distr.	diff.
Trustpilot	0.693	len(text)	0.671	0.618	−0.053	45	0.645	0.622	−0.023
Yelp	0.325	len(text)	0.303	0.195	−0.108	182	0.258	0.220	−0.038

Evaluation result of the sub-experiment.
In individual budget, budgets are applied individually to each text data in dataset, determined by its length, like the conventional word-level approach.
Fixed budget shows the result from the main experiment.

- Overall privacy improvement was more significant (larger difference) in the individual budget approach than in the fixed budget approach

© sebis

Conclusion

RQ1 How can DP be effectively applied at the sentence level within Natural Language Processing, considering the intelligent distribution of privacy budgets for individual words within a sentence?

- Analyze and quantify the importance and informativeness of individual tokens within a text, leveraging linguistic methods to distribute the entire sentence's privacy budget.

RQ2 How can the theoretical concepts of sentence-level privacy with informativeness analysis be translated into an implementable framework?

- Develop a prototype that takes a sentence and the total budget, scores the informativeness of the tokens in the sentence through five methods, and outputs the budget allocated to each token. Apply to existing DP mechanisms.

RQ3 How well does the suggested differential privacy framework protect private data while preserving the utility of the text data?

- The proposed approach shows consistently improved privacy while maintaining usability or with a small loss.

Conclusion – Contribution, Challenges & Future Work

- ✓ Suggesting a new approach to distributing privacy budgets at the sentence level and quantifying informativeness and validating its efficacy.
- ✓ Advancing a practical solution of applying DP in textual data tailored to real-world scenarios with finite privacy budgets

- Quantifying Informativeness of words
  - ⊖ Reliance on statistical methods due to the lack of research on semantic approaches
  - ⏭ Expansion of the prototype with additional scoring methods.
  - ⏭ Adjustment of weights for scoring techniques.
- Budget determination
  - ⊖ It is difficult to estimate the degree of its impact on the data perturbation
  - ⊖ One criterion is used for uniformity of experimental environment settings due to time constraints
  - ⏭ Testing prototypes with varying privacy budgets for insights into effectiveness
  - ⏭ Experimentation with different DP mechanisms and conducting additional tests under various settings and conditions

TUM
sebis

Chaeeun (Joy) Lee

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.0000
chaeeun.joy.lee@tum.de
wwwmatthes.in.tum.de