Driving Context into Text-to-Text Privatization

Stefan Arnold and **Dilara Yesilbas** and **Sven Weinzierl**
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, dilara.yesilbas, sven.weinzierl)@fau.de

Abstract

Metric Differential Privacy enables text-to-text privatization by adding calibrated noise to the vector of a word derived from an embedding space and projecting this noisy vector back to a discrete vocabulary using a nearest neighbor search. Since words are substituted without context, this mechanism is expected to fall short at finding substitutes for words with ambiguous meanings, such as *'bank'*. To account for these ambiguous words, we leverage a sense embedding and incorporate a sense disambiguation step prior to noise injection. We encompass our modification to the privatization mechanism with an estimation of privacy and utility. For word sense disambiguation on the *Words in Context* dataset, we demonstrate a substantial increase in classification accuracy by 6.05%.

1 Introduction

A tension exists between the need to leverage textual data to develop language models and privacy concerns regarding the information conveyed by that data. This is of particular importance because personal information can be recovered from language models (Song and Shmatikov, 2019; Carlini et al., 2020; Pan et al., 2020).

Metric Differential Privacy provides a protection against the disclosure of private information. It has recently been tailored to textual analysis in the form of a text-to-text privatization mechanism (Feyisetan et al., 2020). Building on continuous-valued word embeddings, it relies on the assumption that words close in embedding space serve similar semantic and syntactic roles. This property of embeddings is exploited to replace all words in a text with substitute words given a probability that can be controlled by a noise parameter. A nearest neighbor search is employed to return a substitute word from all words in the embedding space.

A notable deficiency of word embeddings is that they assign a single representation to each word. Depending on its context, an ambiguous word can refer to multiple, potentially unrelated, meanings. Word embeddings are unable to reflect this dynamic nature of words, leading to potentially inappropriate substitutions when used for text-to-text privatization. Clues signaled by inappropriate substitute words may direct a classifier into the opposite direction during downstream tasks. Contextualised word embeddings are an attempt at addressing this limitation by computing dynamic representations for words which can adapt based on context. However, this dynamic behavior makes it virtually impossible to return a substitute word as the nearest neighbor search requires all vectors to be pre-computed and located in the same embedding space.

Sense embeddings represent a middle course between lexical embeddings and contextualized embeddings. By decoupling the static representations of words into multiple representations that capture the meaning of words (covering one representation for each meaning of a word), sense representations enable context-aware text-to-text privatization.

We make the following contributions:

- We replace the word embedding in Feyisetan et al. (2020) with a sense embedding constructed according to Pelevina et al. (2017). To utilize the decoupled senses of words, we further incorporate a word-sense disambiguation prior to the privatization step that discriminates a sense given a sense inventory and a context window.

- We investigate the privacy and utility of substitutions compared to the baseline privatization mechanism without context awareness. Congested by additional representations for each sense of a word, we find that the plausible deniability (acting as our proxy for privacy) is shaped almost identical but allows for smaller noise injection. To demonstrate the utility, we obtain substitutions of identical words paired

Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pages 15–25

in either the same or different contexts. At equivalent levels of privacy, the similarity of substitutions for which their original words belong to the same context show a significantly higher similarity than those of substitutions for which their original words belong to different contexts. Using a set of benchmark tasks from GLUE (Wang et al., 2019), we demonstrate that this difference is an important signal for downstream classification.

2 Preliminaries

2.1 Differential Privacy

Metric Differential Privacy (Chatzikokolakis et al., 2013) is a generalization of differential privacy that originated in the context of location-based privacy, where locations close to a user are assigned with a high probability, while distant locations are given negligible probability. Using word embeddings as a corollary to geo-location coordinates, metric differential privacy has been adopted from location analysis to textual analysis by Feyisetan et al. (2020). This avoids the curse of dimensionality arising from randomized response (Warner, 1965).

We follow the formulation of Xu et al. (2021) for metric differential privacy in the context of textual analysis. Equipped with a discrete vocabulary set $\mathcal{W}$, an embedding function $\phi : \mathcal{W} \rightarrow \mathbb{R}$, where $\mathbb{R}$ represents a high-dimensional embedding space, and a distance function $d : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty)$ satisfying the axioms of a metric (*i.e.*, identity of indiscernibles, symmetry, and triangle inequality), metric differential privacy is defined in terms of the distinguishability level between pairs of words. Formally, a randomized mechanism $\mathcal{M} : \mathcal{W} \rightarrow \mathcal{W}$ satisfies metric differential privacy with respect to the distance metric $d(\cdot)$ if for any $w, w^{'}, \hat{w} \in \mathcal{W}$ the distributions of $\mathcal{M}(w)$ and $\mathcal{M}(w^{'})$ are bounded by Equation 1 for any privacy budget $\varepsilon > 0$:

$$\frac{\mathbb{P}[\mathcal{M}(w) = \hat{w}]}{\mathbb{P}[\mathcal{M}(w') = \hat{w}]} \leq e^{\varepsilon d\{\phi(w), \phi(w^{'})\}}. \quad (1)$$

This probabilistic guarantee ensures that the log-likelihood ratio of observing any word $\hat{w}$ given two words w and w' is bounded by $\varepsilon d\{\phi(w), \phi(w')\}$, providing plausible deniability (Bindschaedler et al., 2017) with respect to all $w \in \mathcal{W}$. We refer to Feyisetan et al. (2020) for a complete proof of privacy. For the mechanism $\mathcal{M}$ to provide plausible deniability, additive noise is in practice sampled from a multivariate distribution such as the *multivariate Laplace distribution* (Feyisetan et al., 2020) or *truncated Gumbel distribution* (Xu et al., 2020a).

We recall that differential privacy requires adjacent datasets that differ in at most one record. Since the distance $d(\cdot)$ captures the notion of closeness between datasets, metric differential privacy instantiates differential privacy when Hamming distance is used, *i.e.*, if $\forall x, x' : d\{\phi(w), \phi(w^{'})\} = 1$. Depending on the distance function $d(\cdot)$, metric differential privacy is therefore generally less restrictive than differential privacy. Intuitively, words that are distant in metric space are easier to distinguish compared words that are in close proximity. Scaling the indistinguishability by a distance $d(\cdot)$ avoids the curse of dimensionality that arises from a large vocabulary $\mathcal{W}$ and allows the mechanism $\mathcal{M}$ to produce similar substitutions $\hat{w}$ for similar w and $w^{'}$. However, this scaling complicates the interpretation of the privacy budget ε, as it changes depending on the metric employed.

Related Work. The multivariate mechanism for text-to-text privatization by Feyisetan et al. (2020) has been extended in orthogonal directions to further improve the utility (Feyisetan et al., 2019; Carvalho et al., 2021) and privacy (Xu et al., 2020b).

Drawing inspiration from Feyisetan et al. (2019), we complement on the line of inquiry dedicated to the enhancement of the utility. By leveraging the curvature of the space at different locations in the Hyperbolic space of Poincaré embeddings (Nickel and Kiela, 2017), their mechanism preserves the hierarchical structure of words during substitution. We persist in the Euclidean space and instead replace the word embedding with a sense embedding to account for the ambiguity of words during substitution. Our results demonstrate that this modification leads to improved performance on downstream tasks while being compatible with prevalent embedding mechanisms.

2.2 Word Embeddings

Since metric differential privacy for text-to-text privatization operates on word embeddings, the merits of privatization are limited by the capabilities of these word embeddings. Starting from sparse vectors suffering from curse of dimensionality, which makes computation and storage infeasible, most research on word embeddings is dedicated to learning dense vectors from corpus-level co-occurrence statistics (Mikolov et al., 2013). To learn these dense vectors, two mirrored approaches have been proposed: continuous bag-of- words and skip-gram. Continuous bag-of- words is trained to predict a word from a fixed window size of context words, whereas skip-gram specifies the probability of observing the context words conditioned on a word within a window. This results in a real-valued vector representation of words that capture interpretable analogical relations between words.

A limitation of these embedding mechanisms is that they conflate all meanings of a word into a single representation, and the most frequent meaning of a word dominates this representation. By conflating all meanings, word embeddings are unable to discriminate ambiguous words. This inability to distinct between ambiguous words is inherited to word substitutions obtained from privatization.

2.3 Sense Embeddings

To address the meaning conflation deficiency of word embeddings, one can represent meanings of words in the form of sense embeddings. Learning sense embeddings has been an active area of research until the emergence of contextual embeddings. We briefly recall some methods to sense representation. Exploiting an unlabeled corpus of text, methods to resolve the meaning conflation deficiency can be divided into three main branches: (1) a staged induction of word senses followed by learning of sense representations, (2) a joint induction of word senses together with learning of sense representations, and (3) retrofitting an existing word embedding by de-conflating word representations into sense representations.

The sense distinctions required to discriminate the meaning of a word are extracted from text corpora by clustering words according to their contexts given a window size. This paradigm is related to word-sense induction. It comes with algorithmic complexity and interpretability problems. Instead of a word-sense induction by clustering, an alternative approach is to derive word senses from predefined sense inventories. This paradigm is related to word-sense disambiguation in which ambiguous words must be assigned a sense from the sense inventory. Exploiting knowledge from pre-defined sense inventories for the initialization of senses allows learning representations that are linked to interpretable sense definitions. Two shortcomings are apparent to learning sense representations using word-sense disambiguation. It is assumed that the sense distinctions intended by the text matches those defined in the sense inventory. Unable to handle words that are not defined in the sense inventory, relying on pre-defined senses hinges on the coverage of the sense inventory.

Staged training of sense embeddings. The training of sense embeddings initially employed a staged approach (Reisinger and Mooney, 2010; Huang et al., 2012; Vu and Parker, 2016). Reisinger and Mooney (2010) constructed sense vectors by clustering sparse vectors corresponding to occurrences of words into a predetermined number of clusters. Clustering is performed by a parametric method that permits controlling the semantic breadth using a per-cluster concentration. Assuming a fixed fixing number of senses for all words, the centroids of the clusters are used as sense vectors and word occurrences are relabeled according to the cluster they belong to. This idea has been extended to dense vectors (Huang et al., 2012).

Instead of inducing senses by clusters, a straightforward method is to disambiguate text corpora as defined by a sense inventory and apply an embedding method on the resulting sense-annotated text (Iacobacci et al., 2015; Flekova and Gurevych, 2016; Ruas et al., 2019). Iacobacci et al. (2015), for instance, use an off-the-shelf disambiguation process to obtain a sense-annotated corpus and directly learn sense representations.

Joint training of sense embeddings. A staged approach to learning sense representations suffers from the limitation that clustering and learning does not take advantage from their inherent similarities. To avoid the issues brought by a two-step clustering, the idea of clustering context vectors has been adapted into the training of word embeddings (Tian et al., 2014; Pina and Johansson, 2014; Neelakantan et al., 2014; Liu et al., 2015b,a; Bartunov et al., 2016; Lee and Chen, 2017; Nguyen et al., 2017). Performing clustering and embedding learning jointly, the intended sense for each word is dynamically selected as the closest sense to the context and weights are updated only for that sense. Assuming a fixed number of senses per word, Tian et al. (2014) introduced an expectation maximization integrated with skip-gram that learns multiple senses weighted by their prior probability. Since words can have a highly dynamic number of senses that range from monosemous words to polysemous words with dozens of associated meanings, this assumption presents a severe limitation. Pina and Johansson (2014) address the varying polysemy problem of sense representation by setting the number of senses of a word as defined by a sense inventory. Deriving the number of senses for each word from a sense inventory, it does not need to create or maintain clusters to discriminate between senses. A better solution would involve dynamic induction of senses from the text corpus. Neelakantan et al. (2014) applies a non-parametric clustering procedure for estimating the granularity of senses for each word. Similar to Tian et al. (2014), it represents the context of a word as the centroid of the vectors of its words but allocates a new sense vector each time the similarity of a context to existing senses is below a certain threshold. By using latent topic modeling to assign topics to each word in a corpus (Liu et al., 2015b,a) and a mixture of weights that reflect different association degrees of each word to multiple senses in the context (Nguyen et al., 2017), words can be discriminated into more general topics.

Retrofitting of word embeddings. Instead of training a word and sense embedding jointly, research exists on refining a word embedding to match semantic constraints (Faruqui et al., 2014; Jauhar et al., 2015; Johansson and Pina, 2015; Rothe and Schütze, 2015; Collier and Pilehvar, 2016). Given a word embedding, Faruqui et al. (2014) propose *retrofitting* as a post-processing step in which words that are connected by a relationship derived from a semantic network are moved closer together in the embedding space. Jauhar et al. (2015) tailored retrofitting towards learning representations for the senses listed in a sense inventory. Using a random walk, Collier and Pilehvar (2016) extracted a set of sense biasing words from an external sense inventory. To de-conflate a word, they add a set of sense embeddings to the same space and push words in the space to the region occupied by its corresponding sense biasing words.

Most retrofitting approaches rely on signals from sense inventories. To transform word embeddings to sense embeddings without external resources, Pelevina et al. (2017) construct a graph by connecting each word to a set of related words. Using ego-network clustering of words, senses are induced as a weighted average of words in each cluster.

2.4 Contextual Embeddings

Although much research has been directed to sense embeddings, the field shifted towards learning contextual embeddings (Peters et al., 2018; Devlin et al., 2019). Rather than pre-computing a static representation for each word, contextualized embeddings dynamically change the representation of a word depending on the context. Harnessing sense signals during the training objective of contextual embeddings has been shown to promote the disambiguation of word meanings (Peters et al., 2019; Huang et al., 2019; Levine et al., 2020; Scarlini et al., 2020). However, the dynamic representations produced by contextual embeddings disqualifies contextual embeddings for privatization as the nearest neighbor search requires that the representations are aligned in a shared embedding space.

3 Methodology

Aiming at context-aware privatization of ambiguous words in texts, we adopt the privatization mechanism of Feyisetan et al. (2020) and replace the word embedding with a sense embedding. The sense embedding is constructed by building and clustering a graph of nearest neighbors based on vector similarities (Pelevina et al., 2017).

Using a context window of size 3 and minimum word frequency of 5, we construct a 300-dimensional word embedding on a dump of `Wikipedia`. We align our vocabulary with words contained in `GloVe`. Our word embedding contains $95,670$ words with words vectors. For each word in the word embedding, we retrieve its 200 nearest neighbors according to the cosine similarity of their word vectors. Once calculated the similarities, we build a graph of word similarities. Assuming that words referring to the same sense tend to be tightly connected, while having fewer connections to words referring to different senses, word senses can be represented by a cluster of words.

A sense inventory is induced from ego-network clustering. The clustering yielded $248,218$ word senses. Each word sense is indexed by a sense identifier. Performing graph clustering of ego-networks is non-parametric. It makes no assumptions about the number of word senses. However, the number and definition of the resulting word senses are not linked to a lexical inventory. Since a word sense is assumed as a composition of words in a cluster, sense vectors are calculated as a weighted pooling of word vectors representing cluster items.

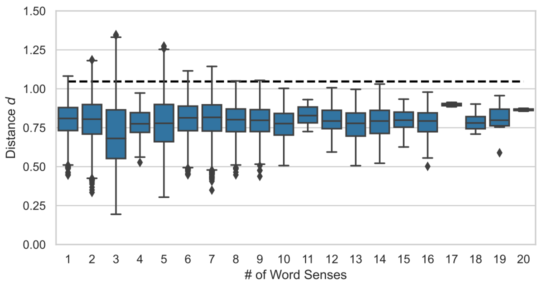


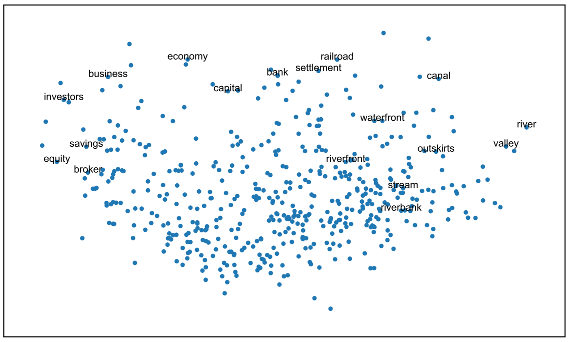
Figure 1: Pairwise Euclidean distances within word senses as a function of the number of distinct senses. The dashed line corresponds to the averaged pairwise distance of word forms in the embedding space.

In Figure 1, we depict the averaged pairwise distances of words as a function of the number of senses. On average, the distance within word senses is considerably lower than the average distance between words in the embedding space (depicted by a dotted line at 1.0550). Since the privatization step is applied directly to the structure of the embedding space, the distance between senses originating from the same word must be taken into account when assessing utility and privacy.

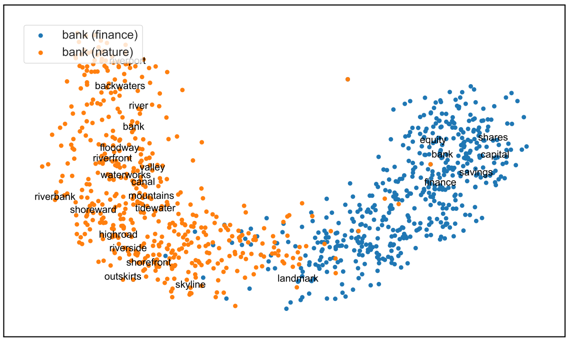
To utilize the sense representations, we incorporate a disambiguation step prior to the privatization. Given a word and its context words, we map the word to a set of its sense vectors according to the sense inventory. The disambiguation strategy is based on similarity between sense and context words: $\operatorname{argmax} \bar{\mathbf{c}}\cdot\mathbf{s}_i / \|\bar{\mathbf{c}}\|\cdot\|\mathbf{s}_i\|$, where $\bar{\mathbf{c}}$ is the mean of the word vectors from the context words. In line with the context size during sense induction, context words for the sense disambiguation are selected within a window of 5. This step is repeated for each word prior to the privatization step.

The privatization step follows a multi-step protocol: We retrieve the sense vector for each disambiguated word. This sense vector is perturbed with noise sampled from a multivariate distribution and its noisy representation is then projected back to the discrete vocabulary space of the sense embedding. As noisy representations are unlikely to exactly represent words in the embedding space, a nearest neighbor approximation is returned. To obtain a private text of word forms, we truncate the sense identifier from the word senses. The result is a privatized text that can be post-processed by word embeddings agnostic to the sense embedding.

To demonstrate the effectiveness of leveraging sense embedding in combination with a disambiguation step prior to the privatization, we privatized the ambiguous word *'bank'* for a total of 500 queries and recorded its substitutions. In half of the queries, the ambiguous word is contained in a text belonging to a geographical context, and in the other half, the ambiguous word is contained in a text belonging to a financial context. The texts are *'to walk by a river **bank** at sunset'* and *to deposit money at a **bank** to earn interest'*. We reduced the dimensionality of the substitute vectors into a two-dimensional space for visualization in Figure 2. We highlight words of the obtained substitutions. We observe that the substitution words returned by lexical privatization stem from both geographical and financial contexts. While substitutions blend between senses during lexical privatization, we discover distinct boundaries between substitute words belonging to contrasting contexts if the words are disambiguated before privatization.



(a) Lexical substitutions for *'bank'*



(b) Contextual substitutions for *'bank'*

Figure 2: Example substitutions associated with a geographical and financial context. A seamless transition in Figure 2(a) compared to distinct regions in Figure 2(b).

4 Experiments

4.1 Privacy Analysis

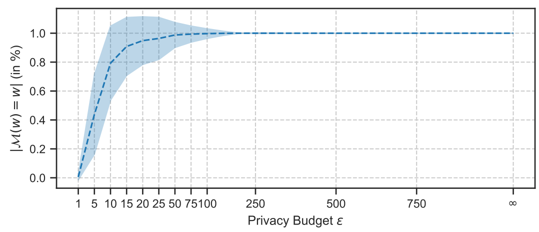
The privacy guarantees in metric differential privacy depend on the deployed metric and the geometric properties of the embedding space. Since retrofitting changes the geometric properties by populating the geometric space of the embedding with word senses that refer to the same word form, we need to recalibrate the plausible deniability (Bindschaedler et al., 2017). We record the following statistics as proxies for the plausible deniability. We note that these proxy statistics have been used in previous studies to characterize the plausible deniability of multivariate mechanisms (Feyisetan et al., 2019, 2020; Xu et al., 2020b, 2021).

- $N_w = \mathbb{P}\{M(w) = w\}$ measures the probability that a word is not substituted by the mechanism. This is approximated by counting the number of occurrences in which a word w is substituted by the same word after running the mechanism for 100 times.

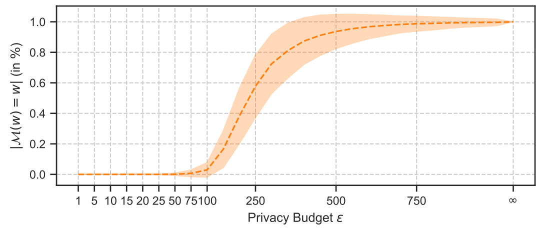
- $S_w = |\mathbb{P}\{M(w) = w^{'}\}|$ measures the effective support in terms of the number of distinct substitutions produced for a word from the mechanism. This is approximated by the cardinality of the set of words $w^{'}$ after running the mechanism for 100 times.

Since the noise in the multivariate Laplace mechanism is scaled by $1/\epsilon$, we can make a connection between the proxy statistics and the privacy budget ϵ. A smaller ϵ corresponds to more stringent privacy guarantees by adding more noise to the word embedding. More noise leads to fewer unperturbed words (lower N_w) and more diverse outputs for each word (higher S_w). By contrast, a higher ϵ leads to less substitutions (higher N_w) and a narrow set of distinct words (lower S_w). From a distributional perspective, it follows that N_w (S_w) should be positively (negatively) skewed to afford reasonable privacy guarantees.

In Figures 3 and 4, we present the averaged values of N_w and S_w over 100 independent queries from the corpus of `WikiText` (Merity et al., 2016) for a discrete set of privacy budgets $\varepsilon = \{1, 5, 10, 15, 25, 50, 100, 250, 500, \infty\}$. While lower values of ε are desirable in terms of privacy, plausible deniability is assured unless N_w (S_w) exceeds (falls below) 0.5. The plots thus serve as a visual guidance for comparing (and selecting) the privacy budget ε. The curve of the privacy proxies as function of the privacy budget is shaped identical for word and sense embeddings, except that using a sense embedding stretches the allocatable privacy budget by an order of magnitude. We attribute this shape to the congestion of the embedding space with substitution candidates, even at low levels of noise.



(a) Lexical N_w



(b) Contextual N_w

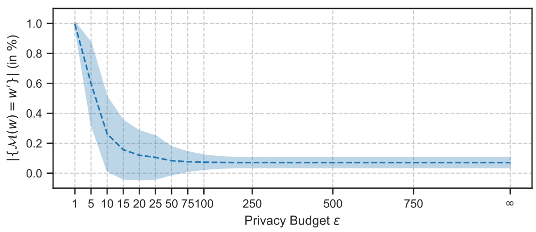
Figure 3: N_w refers to the number of substitute words that are *identical* to a queried sensitive word. The shift in the curve suggests that higher privacy budgets are legitimate before there is a risk that words will not be replaced by substitutions.

For our utility experiments, we set the privacy budget for each mechanism so that .90 quantile of words is plausible deniable. To calculate the .90 quantile, we interpolated the scores for N_w (S_w) and selected the privacy budget ε so that N_w (S_w) does not exceed (fall below) 0.5. A plausible deniability for only a quantile of words was also assumed in a prior study by Xu et al. (2020b).

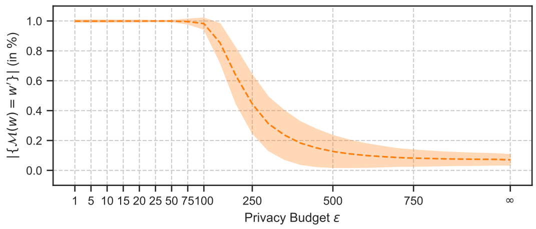
4.2 Utility Analysis

To analyze the utility of privatization with context awareness, we use the standard datasets for evaluating word similarity. The datasets include `WordSim-353` (Agirre et al., 2009), `SimLex-999` (Hill et al., 2015), and `SWCS` (Huang et al., 2012). Common to all these datasets is that similarity ratings are given to pairs of words. While `WordSim-353` and `SimLex-999` provide pairs of words in isolation, `SWCS` provides a context for each word that triggers a specific meaning, making it very suitable for the evaluation of context-aware privatization. All experiments are conducted while ensuring plausible deniability for .90-quantile of words.

We query each pair of words (w_i, w_j) for 25 times by each privacy mechanism and record their



(a) Lexical S_w



(b) Contextual S_w

Figure 4: S_w refers to the number of substitute words that are *unique* from a queried sensitive word. The shift in the curve suggests that higher privacy budgets are legitimate before the effective support of substitution candidates violates plausible deniability.

	(w_i, w_j)	**Words**	**Senses**
WordSim-353	0.5849	0.1353	0.1478
SimLex-999	0.2978	0.0696	0.0841
SCWS	0.5183	0.1911	0.2358

Table 1: Datasets for measuring the similarity between words. Similarity measured after substitution. Scores denote the correlation compared to annotations.

similarity after privatization. We use the cosine distance as our similarity measure. The results capture $\hat{\mathbf{w}}_i \cdot \hat{\mathbf{w}}_j / \|\hat{\mathbf{w}}_i\| \cdot \|\hat{\mathbf{w}}_j\|$. Once queried, we correlate the measured similarity against the similarity annotations. We present the results in Table 1. Without a context provided to discriminate a word, the privatisation using sense embeddings generalizes to privatisation using word embeddings. This can be seen by the almost identical correlation coefficients for `WordSim-353` and `SimLex-999`. The correlation of the sense embedding surpassing those for the word embedding on `SWCS` indicates that the information provided by the disambiguation step helps in finding more appropriate substitutions.

We further benchmark our mechanism in combination with a `BERT` model for downstream classification. We employ the words in context (Pilehvar and Camacho-Collados, 2019) dataset. It is composed of $5,428$ text-pairs for training and 638 text-pairs for validation. Framed as a binary classification task, the goal of words in context is to identify if the occurrences of a word for which two contexts are provided correspond to the same intended meaning. Each of context is designed to trigger a specific meaning. Note that the dataset is balanced, hence, a context-insensitive embedding would perform similarly to a random baseline.

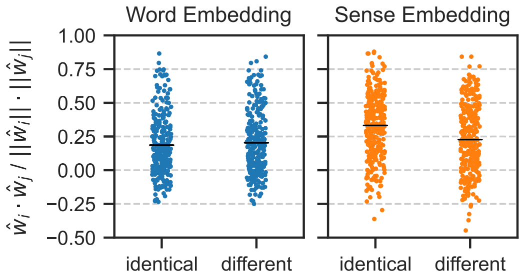


Figure 5: Cosine similarity of word pairs after substitution. The vertical line represents the average similarity.

Without privacy guarantees, `BERT` peaks at an accuracy score of 0.6887. The training using the privatized data mimics the training without privatization. After privatizing the training data using word embeddings, `BERT` scores 0.6006. Leveraging sense embeddings, we boost the accuracy to 0.6423. This narrows the gap in accuracy by 6.05%. All scores are calculated as an average over three independent trials for each privatization mechanism.

To provide an explanation for the substantial improvement, we queried each record in the words in context dataset for 25 times and recorded the cosine similarity between the word pairs after substitution. Since we are only interested in the instances a substitution occurs, we removed cases in which the similarity between substitutions is one. We expect that the similarity between $\hat{w}_i$ and $\hat{w}_j$ obtained from the privatization step is higher when w_i and w_j belong to the same context and lower when different contexts are intended. Whether the words are from an identical context or different contexts is directly derived from annotations. For a transparent comparison, we measure the similarity using `GloVe` representations of their corresponding substitutions. We present the results in Figure 5, separated by word and sense embedding.

The representations of substitutions obtained by a word embedding convey no clues about the intended contexts the word belongs to. This can be argued by an average similarity that is almost identical at values of 0.1860 and 0.2035. Compared to

	Level of Privacy	Classification CoLA (MCC)	SST2 (ACC)	Textual Similarity QQP (ACC)	MRPC (ACC)	STSB (SCC)	Textual Entailment MNLI (ACC)	QNLI (ACC)	RTE (ACC)	Avg. -
BERT	-	0.5792	0.9243	0.8879	0.8329	0.8854	0.8229	0.8912	0.6927	0.8146
Words	p=0.9	0.0000	0.7614	0.6883	**0.6059**	0.5619	0.5270	0.6145	0.5342	0.5367
	p=0.5	0.0416	0.8518	0.7858	0.6123	0.5907	0.7001	0.7893	0.5880	0.6200
Senses	$p = 0.9$	0.0000	**0.8669**	**0.7715**	0.5910	**0.6197**	**0.6750**	**0.7446**	**0.5834**	**0.6065**
	$p = 0.5$	0.0655	0.8862	0.8215	0.6322	0.6442	0.7417	0.8180	0.6070	0.6520

Table 2: Results on a subset of GLUE (Wang et al., 2019). We report Matthews correlation for the CoLA dataset, Spearman correlation for the STSB dataset, and the accuracy score for all remaining datasets. The level of privacy increases with the quantile of words that are provable plausible deniable. $p = .90$ denotes an (almost) worst-case scenario. $p = .50$ denotes an average-case scenario. Bold font indicates the best result from three independent trials.

the similarity of lexical representations, the average similarity of substitutions within the same context is 0.3118 and 0.2272 for words that originate from different contexts. This distinguishability signals whether words are paired in identical or different contexts, which indicates an awareness of the context during privatization.

We expect the awareness of the meaning of words to carry over to downstream tasks. To thoroughly evaluate whether context-awareness during privatization translates into better performance on downstream tasks, we conduct experiments on a set of classification tasks in the text domain. We use the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). GLUE is a collection of diverse language understanding tasks. The benchmark involves classification of ordinary text and text pairs for similarity and entailment. Apart from CoLA (Warstadt et al., 2019), which requires high level of syntactic reasoning, all other tasks are based on semantic reasoning.

We summarize the results on a subset of GLUE obtained by fine-tuning a pre-trained BERT (Devlin et al., 2019) in Table 2. We report the scores once for word embeddings and once for sense embeddings. Using sense embeddings as opposed to word embedding, the average performance increases from 0.5367 to 0.6065. This result confirms our expectation that context awareness during privatization translates into better performances on downstream tasks.

5 Conclusion

We redesigned the multivariate mechanism of metric differential privacy in the text domain to account for word meaning during privatization. We accomplished this by replacing the word embedding with a sense embedding and incorporating a sense disambiguation step prior to the noise injection.

Despite the congestion of the embedding space with senses that stem from the same word form, we experimentally demonstrated that our modification follows the privacy formalization of Feyisetan et al. (2020). Once we recalibrated the privacy budget to ensure plausible deniability, we measured the capability of our mechanism to capture the word meaning. By calculating the similarity of pairs of words in a context that triggers the meaning of each word, we observe that the similarity score for substitutions is consistently higher when both words appear in the same context, and lower when both words appear in different contexts.

With the confirmation that our mechanism captures word meaning, we were interested in whether the benefits of contextual substitutions translates into superior performance in downstream classification tasks. The results on a set of benchmark datasets demonstrated a substantial boost in generalization performance for tasks that rely on semantic reasoning rather than syntactic reasoning.

Limitations. Our modification utilizes sense embeddings. Since the senses were not mapped to an external inventory, the senses cannot be interpreted. Apart from the lack of interpretability, sense embeddings are superseded by contextual embeddings derived from transformer models with sense awareness (Huang et al., 2019; Levine et al., 2020; Scarlini et al., 2020). While sense embeddings and contextual embeddings are not mutually exclusive, it is necessary to alternate between them for the purpose of privatization and optimization.

Acknowledgment

We gratefully acknowledge that this research was supported in part by the *German Federal Ministry of Education and Research* through the *Software Campus* (ref. *01IS17045*).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*, pages 130–138. PMLR.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021. Tem: High utility metric differential privacy on text. *arXiv preprint arXiv:2107.07928*.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

Nigel Collier and Mohammad Pilehvar. 2016. Deconflated semantic representations. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105.

Sujay Kumar Jauhar, Chris Dyer, and Eduard H Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *HLT-NAACL*, pages 683–693.

Richard Johansson and Luis Nieto Pina. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433.

Guang-He Lee and Yun-Nung Chen. 2017. Muse: Modularizing unsupervised sense embeddings. *arXiv preprint arXiv:1704.04601*.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 4656–4667, Online. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015a. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Twenty-fourth international joint conference on artificial intelligence*.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015b. Topical word embeddings. In *Twenty-ninth AAAI conference on artificial intelligence*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2017. A mixture model for learning multi-sense word embeddings. *arXiv preprint arXiv:1706.05111*.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2017. Making sense of word embeddings. *arXiv preprint arXiv:1708.03390*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Nieto Pina and Richard Johansson. 2014. A simple and efficient method to generate word sense representations. *arXiv preprint arXiv:1412.6045*.

Joseph Reisinger and Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing, China. Association for Computational Linguistics.

Terry Ruas, William Grosky, and Akiko Aizawa. 2019. Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications*, 136:288–303.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8758–8765.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160.

Thuy Vu and D Stott Parker. 2016. k-embeddings: Learning conceptual embeddings for words using context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1262–1267.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2020a. Differentially private adversarial robustness through randomized perturbations. *arXiv preprint arXiv:2009.12718*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020b. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a utilitarian approach to privacy preserving text generation. *arXiv preprint arXiv:2104.11838*.