Privacy guarantees for de-identifying text transformations

David Adelani, Ali Davody, Thomas Kleinbauer, Dietrich Klakow

Privacy Guarantees for De-identifying Text Transformations

David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow

Spoken Language Systems Group, Saarland Informatics Campus, Saarland University, Germany

{didelani|kleiba|adavody|dietrich.klakow}@lsv.uni-saarland.de

Abstract

Machine Learning approaches to Natural Language Processing tasks benefit from a comprehensive collection of real-life user data. At the same time, there is a clear need for protecting the privacy of the users whose data is collected and processed. For text collections, such as, e.g., transcripts of voice interactions or patient records, replacing sensitive parts with benign alternatives can provide de-identification. However, how much privacy is actually guaranteed by such text transformations, and are the resulting texts still useful for machine learning?

In this paper, we derive formal privacy guarantees for general text transformation-based de-identification methods on the basis of *Differential Privacy*.

We also measure the effect that different ways of masking private information in dialog transcripts have on a subsequent machine learning task. To this end, we formulate different masking strategies and compare their privacy-utility trade-offs. In particular, we compare a simple *redact* approach with more sophisticated *word-by-word* replacement using deep learning models on multiple natural language understanding tasks like named entity recognition, intent detection, and dialog act classification. We find that only word-by-word replacement is robust against performance drops in various tasks.

Index Terms: Differential privacy, Spoken language understanding, Named entity recognition, Intent detection.

1. Introduction

Machine learning approaches, in particular Deep Learning, dominate many areas of Natural Language Processing (NLP). To reach peak performance, they require large data sets to train models. It is thus common to continuously collect user data after a model has been deployed in order to augment existing training data. This practice raises a clear need for protecting the privacy of the users whose data are collected. For instance, commercial providers of voice assistants have been criticized for recording and transcribing conversations of their users[1]. But other domains are affected as well, for instance, patients' health records in medical applications.

For text collections, one way to respect the users' privacy is to sanitize each document through a de-identification process before adding it to a data collection. De-identification requires to either delete all sensitive information in a text, or to replace it with benign surrogates. Arguably, strict deletion is the less desirable option because without an indication of where the edit was made, texts can become impossible to understand or, perhaps worse, change their meaning. To illustrate, consider the following example from the medical domain, where the names of specific medications are considered sensitive information:

(1) Besides *warfarin*, the patient is not taking any medication.

(2) Besides, the patient is not taking any medication .

A more commonly used alternative to deletion is *redaction*, where relevant text portions are blackened rather than deleted, but this can still impact readability. For instance, when dates, times, and locations are considered sensitive with respect to a person's whereabouts, a sentence such as (3) would be rather useless in a text corpus in its redacted form (4):

(3) How about *Rick's Café* around *noon* on *the 15th* ?

(4) How about [redacted] around [redacted] on [redacted] ?

This example illustrates that de-identification can imply a trade-off between *privacy* and *utility*. In order to gauge the latter, Tang et al. measure the impact of three alternative methods for masking sensitive words on a subsequent machine learning task [1]. The methods consist of two different ways of replacing words with other, randomly selected words from the same category, and of a method for replacing words with a specific category marker. Applying these strategies to the previous example could lead e.g. to the sentences in (5) and (6) respectively:

(5) How about *London* around *4 o'clock* on *the 3rd of May* ?

(6) How about *<LOCATION>* around *<TIME>* on *<DATE>*?

However, Tang et al. [1] do not give any formal privacy guarantees for their methods, making it difficult to judge the privacy-utility trade-off.

In this paper, we fill this gap by deriving formal privacy guarantees for general text transformation-based de-identification methods on the basis of *Differential Privacy*, a well-established framework for quantifying privacy leakage [2]. In addition, we show the impact of five different text transformation strategies on three common NLP tasks, when the transformed texts are used as training data for machine learning approaches. Unlike Tang et al., we perform our experiments on six different corpora to gain more balanced evidence.

2. Related Work

Text de-identification, also known as sanitization, is well established in highly sensitive domains, such as e.g., for patient health records [3]. A large number of de-identification methods have been suggested in that area in the past e.g. [4, 5, 6]. While such methods are important for a number of domains, we focus here on the privacy issues associated with cloud-based dialog systems, such as e.g. [7], which are generally acknowledged (e.g. [8]) but have not yet received wide-spread attention.

In contrast, the necessity for protecting private information has long been realized in the data mining community [8]. Our task is, however, quite different from data mining. For instance, data mining transformations oftentimes pay special attention to

This research has received funding by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 3081705 – COMPRISE (http://www.compriseh2020.eu/)

[1] https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio

the preservation of certain statistical properties of the underlying data, which is not a primary concern in our work, thus allowing us to explore simpler approaches.

Probably the closest work to ours is [1] where the impact of data sanitization has been investigated on Automatic Speech Recognition (ASR) and call classification tasks. It has been shown there that the spoken dialog system trained on sanitized data achieves a comparable accuracy. In contrast, we apply text replacement to Natural Language Understanding (NLU) tasks and show that some replacement strategies can have a large destructive effect on the performance of the model. We verify this by applying a state-of-the-art deep learning model to train the NLU tasks by fine-tuning BERT [9] embeddings on three tasks that include named entity recognition, intent detection and dialog act classification.

Recently, Carrell et al. [10] proposed an attack to leak sensitive information in a transformed text, however, this attack only works on a small dataset. The attack does not scale to large datasets because it requires the attacker to perform annotation of private tokens, which is costly and tedious.

3. Privacy

A general framework for protecting privacy is *Differential Privacy* (DP) introduced in [11]. DP quantifies to what extent privacy in statistical queries is preserved while extracting useful information from a dataset and has received increasing attention recently as a rigorous privacy methodology. In this section, we clarify the connection between DP and text replacement methods but first provide some technical background on general DP.

Let $\mathcal{D}$ be the set of all possible datasets for a given domain of data points. A key concept in DP is neighboring datasets. We call two datasets $D_1, D_2 \in \mathcal{D}$ neighboring if they are the same except for one data point. For example, D_1 and D_2 could be two text corpora which differ only in one single word. The intuition behind differential privacy, as defined below, is a guarantee that a randomized algorithm behaves similarly on similar input datasets to a point where the output of the algorithm does not allow to infer which dataset was used with any relevant degree of certainty. Therefore, an attacker cannot tell whether the aforementioned data point is contained in the algorithm's dataset or not.

Definition. *(Differential Privacy). A randomized algorithm $\mathcal{M}$ is (ε, δ) private with domain $\mathcal{D}$ if for all measurable sets $S \in$ Range$(\mathcal{M})$ and for all neighboring datasets D_1 and D_2 differing in at most one data point, we have*

$$\Pr[\mathcal{M}(D_1) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D_2) \in S] + \delta \quad (1)$$

Intuitively, a (ε, δ) differential private mechanism guarantees that the absolute value of privacy leakage will be bounded by ε with probability at least $1 - \delta$ for adjacent datasets. The higher the value of ε, the higher the chance of data re-identification.

Algorithm: Probabilistic Text De-identification

Input: dataset $\mathcal{D}$, token replacement policy π, probability parameter p.
for t' *in sensitive data* **do**
  $r \sim U(0,1)$ **if** $r \leq p$ **then**
    replace t' with $t \sim \pi(t|t')$
  end
end

To define a general algorithm for de-identification, let $\mathcal{T}$ denote the vocabulary of private tokens and consider a token replacement policy $\pi : \mathcal{T} \longrightarrow \mathcal{T}$, where $\pi(t|t')$ is the probability of replacing t' in the original text with t. We introduce a parameter p to model the probability that a token gets replaced:

Lemma. *If token replacement policy π in the algorithm is independent of the token to replace, i.e. $\pi(t|t') = \pi(t)$, the algorithm is $(\varepsilon, 0)$ differentially private with:*

$$\varepsilon = \max_t \log \frac{1 - p + p\,\pi(t)}{p\,\pi(t)}. \quad (2)$$

To prove it, we consider two neighboring datasets D_1 and D_2, which are the same except in one token. In other words, D_2 can be obtained from D_1 by replacing a token t_1 in D_1 with t_2. Using this notation, we may compute the privacy loss as follows. Let

$$\varepsilon = \log \frac{\Pr[t \in \mathcal{M}(D_1)]}{\Pr[t \in \mathcal{M}(D_2)]} \quad (3)$$

where $\mathcal{M}(D)$ is the dataset obtained by applying the de-identification algorithm to the original dataset D, and t is the observed token in the resulting text. If t_1 and t_2 are not equal to t (i.e., t_1 and t_2 are replaced by t), we have :

$$\frac{\Pr[t \in \mathcal{M}(D_1)]}{\Pr[t \in \mathcal{M}(D_2)]} = \frac{p\,\pi(t|t_1)}{p\,\pi(t|t_2)} = 1 \quad (4)$$

where we have used this fact that replacement policies are independent of the original tokens. On the other hand, if t is equal to t_1, we arrive at the following expression for the privacy loss:

$$\frac{\Pr[t \in \mathcal{M}(D_1)]}{\Pr[t \in \mathcal{M}(D_2)]} = \frac{1 - p + p\,\pi(t)}{p\,\pi(t)}. \quad (5)$$

We get the inverse of this expression in the opposite case $t = t_2$. The overall privacy bound is given by the maximum of (4) and (5) over the private tokens as stated in (2). □

The algorithm is a variant of *randomized response* [12] whose connection to differential privacy has been studied before (e.g. [2, 12, 13]), although not in the context of text de-identification. The probability parameter p gives data curators fine-grained control over the privacy-utility trade-off: an *ideal* text replacement, corresponding to $p = 1$, has zero privacy loss ($\epsilon = 0$) but in cases where the replacement noise harms the performance of models trained on the resulting data too much, the curator might choose to use a lower probability p if reduced privacy is deemed acceptable. Our lemma allows to quantify this effect and compare different de-identification options.

In practice, $p = 1$ cannot be achieved very easily if the sensitive tokens are identified automatically as in e.g. [1]'s and our own experiments below. Instead, the *recall* value of the employed identification method defines an upper bound for p, e.g., a recall value of 0.8 implies that an expected 20% of the sensitive tokens will not be replaced. As p approaches 0, the ϵ value approaches infinity, meaning that no privacy is provided.

In order to interpret our result, we consider the case where an attacker gets hold of the fully transformed data set. The level of privacy expressed by the lemma refer to the possibility of reversing the replacement in order to reconstruct the source from a transformed sentence, which is difficult when the privacy loss is small. Context information might be helpful, but in general, original tokens can only be guessed according to their prior probabilities which we assume to be uniform in this paper. However, the algorithm allows for certain sentences to appear in the output untransformed, either because of the value of the randomized response value r, or when the randomly chosen replacement token happens to be identical to the source token. The privacy guarantee given by our lemma arises from

Table 1: *Examples of the replacement strategies, using color codes for PER, LOC, ORG, and TIME.*

Replacement strategy	Transformed text
No Replacement	Hi Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving by six pm
Redact	Hi Mister IIIII, the IIIII flight from IIIII to IIIII is leaving by IIIII
Typed-Placeholder	Hi Mister PER, the ORG flight from LOC to LOC is leaving by TIME
Named-Placeholder	Hi Mister Smith, the SAP flight from London to London is leaving by afternoon
Word by word	Hi Mister John, the BOSCH flight from New Boston to Berlin is leaving by eleven morning
Full entity	Hi Mister John, the BOSCH flight from New York to Berlin is leaving by twelve pm

the fact that transformed and untransformed sentences are not obviously distinguishable. In fact, the DP parameter, ϵ can be seen as a measure of the certainty with which an attacker can judge whether a sentence from the output was actually part of the source text.

Referring to a specific instance of the above algorithm, i.e. a fixed choice for p and π (called a text replacement *strategy*), with tokens being either single words or multi-word expressions, we examine some straight-forward replacement strategies and the level of privacy they present in the light of our results.

Redact Here, the private tokens are replaced with a non-word placeholder that is typically not part of the vocabulary of the source text e.g. `IIIII`. Hence, we only fall into case (4) above, implying $\epsilon = 0$ under the interpretation outlined above: an attacker can decide with certainty which of the tokens were part of the original text but cannot infer the replaced tokens.

Typed placeholder (aka value-class membership [1]) This is akin to using private category markers like LOCATION as the replacement token. This is a strategy similar to redaction, providing the same level of privacy. However, it provides additional information about a replaced token's category and might thus be more useful than redaction for certain NLP tasks.

Named placeholder A fixed category exemplar is used to replace all private tokens of that category [14], e.g, all locations are replaced by "London". This strategy makes it slightly more difficult to judge which sentence was transformed and which was not, i.e. $\epsilon > 0$. But for all instances that differ from the exemplar, it is clear that they must have been part of the source.

Word-by-word replacement We can distinguish between *value distortion* [1] if the replacement tokens are from an external source, and *value dissociation* [1] when the surrogate tokens are from the same corpus. The latter keeps the distribution of tokens in the resulting document unchanged, which might be relevant for some tasks. Both variants make it hard to identify untransformed sentences, which is reflected in lower ϵ values.

Full entity replacement Text coherence could be improved if source tokens were consistently replaced by the same surrogates. However, this case is not supported by our lemma where we require $\pi(t|t') = \pi(t)$. Another downside of the word-by-word strategy is that multi-word expressions could lead to nonsensical replacements, e.g. "Frankfurt Airport" could be transformed to "New Francisco". A variant is thus to replace full entities instead of single words. In terms of what can be captured by our lemma, this does not lead to more privacy, but the expected gain in coherence might benefit downstream tasks.

An example for each of these replacement strategies is given in Table 1. Besides discussing privacy aspects, we have speculated on the differences of the strategies on subsequent applications. In order to verify these considerations, we now measure the impact of the different replacement strategies empirically.

4. Utility

We experiment with three common NLP tasks, Named Entity Recognition (NER), Intent Detection (ID), and Dialog Act Classification (DAC), across six different datasets (see Table 3). The variety in datasets is important since what is considered sensitive information is typically domain-dependent. Here, we consider as private: (1) the identity of one or both speakers, (2) organizations, such as e.g., company names, etc. (3) The locations or addresses (4) The dates and times. This private information coincides with typical *named entities (NEs)* and *slot classes* in dialog datasets such as PER (personal names), ORG (organization), LOC (location), DATE and TIME.

4.1. Datasets

The **VERBMOBIL** corpus is a large collection of spontaneous telephone conversations [15]. In each conversation, two speakers negotiate the details of a business meeting. The corpus contains English, German, and Japanese conversations, however, we only use English portion of the corpus for our experiments. The VERBMOBIL corpus does not come pre-annotated with NE classes. About 20% of the VERBMOBIL corpus was thus annotated via crowd sourcing. The remaining 80% of the corpus was annotated automatically using `spaCy`[2] and post-corrected manually.

The **ATIS** [16] corpus is a popular dataset for slot filling and intent detection tasks in the Air Travel Information Services domain. For the text transformation experiments, we map the provided slot labels to the aforementioned named entity categories.

SNIPS is another popular benchmark dataset for slot filling and intent detection task by SNIPS.AI [17]. The dataset consists of seven intents from different domains such as *"AddToPlaylist"*, *"BookRestaurant"*, *"GetWeather"*, *"RateBook"*.

FB en-TOD is a multilingual slot and intent classification dataset recently released by Facebook [18]. It consists of utterances from three languages (English, Spanish, and Thai) and three domains (Alarm, Reminder, and Weather). In this paper, we only use the English dataset.

MS Taxi and **MS Restaurant** are two out of three dialog challenge datasets released by Microsoft at the SLT 2018 workshop [19] for taxi bookings and restaurant reservations with 19 and 29 slot types respectively and 11 dialog acts. The number of classes for the Taxi and Restaurant datasets are 18 and 24 respectively after removing classes with less than 40 utterances.

4.2. Experiments

For all comparison experiments, we first run a baseline experiment using the original datasets. Then, we apply the respective privacy strategies to the training data before fine-tuning a BERT model for token/sentence classification. We then compare the performance of the resulting models with the baseline with respect to the (untransformed) test set. The BERT classification model involves *fine-tuning* the pre-trained BERT embeddings on the training data with an additional linear layer whose weights are randomly initialized. We trained all the parameters of the model end-to-end, including the linear layer.

[2] `https://spacy.io`

Table 2: *Evaluation of the different token replacement strategies on the 6 datasets comprising of 3 tasks: NER, ID, DAC. Average performance computed from ten runs. The best Accuracy/F1-score in each class/task are in* ***bold*** *and the best text transformation result have asterisk (*). The replacement strategies use ground-truth annotations for the identification of sensitive tokens, i.e.* $p = 1, \varepsilon = 0$.

Replacement strategy	VerbMobil NER F1-score	ATIS ID Accuracy	SNIPS ID Accuracy	en-TOD ID Accuracy	Restaurant DAC Accuracy	Taxi DAC Accuracy
No replacement	$\mathbf{88.3 \pm 0.2}$	$\mathbf{98.4 \pm 0.2}$	$\mathbf{98.0 \pm 0.2}$	$\mathbf{99.4 \pm 0.0}$	$\mathbf{78.9 \pm 0.1}$	$\mathbf{90.0 \pm 0.1}$
Redact	0.2 ± 0.2	94.8 ± 0.2	89.7 ± 0.8	97.4 ± 0.6	75.9 ± 0.3	88.1 ± 0.2
Typed-Placeholder	0.0 ± 0.0	95.7 ± 0.3	54.1 ± 3.8	97.2 ± 0.7	76.5 ± 0.2	87.9 ± 0.5
Named Placeholder	13.5 ± 1.4	95.9 ± 0.3	76.2 ± 2.9	98.2 ± 0.1	77.3 ± 0.2	89.3 ± 0.1
Word-by-Word	72.6 ± 0.3	$\mathbf{98.6 \pm 0.2}^*$	$97.5 \pm 0.3^*$	$99.2 \pm 0.1^*$	78.4 ± 0.2	$89.9 \pm 0.2^*$
Full Entity	$85.9 \pm 0.3^*$	$\mathbf{98.5 \pm 0.2}^*$	$97.4 \pm 0.3^*$	$99.2 \pm 0.1^*$	$78.5 \pm 0.1^*$	$89.9 \pm 0.1^*$

Table 3: *Dataset summary for three different tasks: Named Entity Recognition (NER), Intent Detection (ID), and Dialog Act Classification (DAC)*

Dataset	Task	Private Tokens	Classes	Train / Val. / Test Sentences
VerbMobil	NER	5 NEs	6	19K / 2848 / 5230
ATIS	ID	21 slots	21	4478 / 500 / 893
SNIPS	ID	39 slots	7	13K / 700 / 700
FB en-TOD	ID	15 slots	12	30K / 4181 / 8621
MS Restaurant	DAC	21 slots	24	20K / 2936 / 5859
MS Taxi	DAC	10 slots	18	16K / 2273 / 4597

For the implementation, we fine-tuned BERT on the various tasks using the *simpletransformers*[3] based on the *transformers* library of HuggingFace [20]. The hyper-parameters of the model are 768-dimensional embedding layer (for bert-base-cased model), batch size of 8 for NER and 16 for other tasks, maximum learning rate of 0.00005, maximum sequence length is 128 for NER and 64 for other tasks. The maximum number of epochs for all experiments is 3.

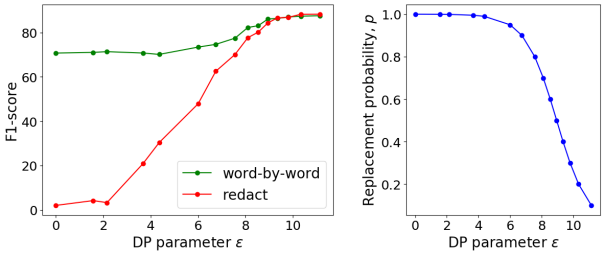

Figure 1: *Connection between DP (privacy), Replacement probability p, and F1-score (performance) on* VERBMOBIL.

4.3. Results

We show that the performance of the redact and word-by-word replacement strategies can be improved by tuning the parameter p described in Section 3. For example, by setting $p = 0.9$ (i.e. $\epsilon = 6.75$), we can improve the F1-score by around 4% for word-by-word and 60% for redact as shown in Figure 1, demonstrating the ability to control the privacy-utility trade-off. In the word-by-word replacement experiments, we replace a NE word t' by another word t of the same entity class based on their relative frequency distribution $\pi(t)$ in the corpus.

The baseline to which we compare all other experiments is simply trained on the original training set, i.e., without removing any private information. On the test set, the resulting model yields a prediction F1-score/Accuracy of 88.3% (NER), 98.4% (ATIS intent) and 98.0% (SNIPS intent), 99.4% (en-TOD intent), 78.9% (Restaurant DAC) and 90.0% (Taxi DAC).

Table 2 shows the result for the different text transformation strategies. Replacing private tokens using *redact*, *typed placeholder* and *named placeholder* strategies generally gave a worse result than the word-by-word replacement. For NER, we observe a substantial drop in performance for redact and placeholder approaches because the model overfits on the replacement tokens which are expected to be absent in the test set. On the other hand, the drop is minimal for intent and dialog act classification tasks around $(2 - 4\%)$ similar to the observation in [1], except for the SNIPS dataset with much larger reduction in performance of $8 - 44\%$ depending on the placeholder strategy. This shows that these transformation strategies are generally not suitable for training NLU systems.

For the *word-by-word* replacement, we observe a drop of 15% in F1-score when we replace all words labeled as named entities with tokens of the same-type. For NER, we find that "TIME", "ORG" and "DATE" are most affected by the word-by-word replacement in terms of drop in F1-score because many of them are multi-word expressions. Thus, the three named entities gain the most by full-entity replacement. On the other-hand, the drop is very small ($< 1\%$) for other intent and dialog act classification.

Table 1 illustrates an example of the full-entity replacement (e.g "Frankfurt Airport" is replaced by "New York"). This approach gives the best performance out of all the transformation strategies with only 2.4% drop for NER. Interestingly, there is no significant difference between its performance and the baseline on the intent and dialog act classification tasks across the datasets. In summary, the text obtained using the word-by-word or full-entity text transformation are more suitable for training NLU systems while protecting the privacy of users.

5. Conclusion

Replacing sensitive tokens with benign alternatives is a common method for de-identifying text documents. We prove that privacy guarantees for a formalized version of this process can be expressed in terms of Differential Privacy. Our approach includes two parameters, p and π that allow different replacement strategies to be expressed as instances of the same algorithm. The respective DP-ϵ value follows from the choices for p and π, permitting a comparison of different replacement strategies with respect to their privacy implications.

User privacy is juxtaposed by the performance impact that a text transformation has on subsequent machine learning tasks. We experiment with three different NLP tasks across six different datasets and find that both word-by-word and full entity replacement strategies are robust against performance drops across all examined tasks.

[3] https://github.com/ThilinaRajapakse/simpletransformers

6. References

[1] M. J. Tang, D. Z. Hakkani-Tür, and A. GokhanTur, "Preserving privacy in spoken language databases," in . *In Proc. of the International Workshop on Privacy and Security Issues in Data Mining, ECML/PKDD*, 2004.

[2] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*, ser. Foundations and Trends in Theoretical Computer Science. Now Publishers Inc., August 11, 2014.

[3] Özlem Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, September 2007.

[4] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research," *BMC Medical Research Methodology*, vol. 10, no. 70, 2010, https://doi.org/10.1186/1471-2288-10-70.

[5] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, June 2017.

[6] K. Khin, P. Burckhardt, and R. Padman, "A deep learning architecture for de-identification of patient notes: Implementation and evaluation," in *Proceedings of the Workshop on Information Technologies and Systems (WITS)*, San Jose, CA, USA, December 16-18 2018.

[7] H. Chen, X. Liu, D. Yin, and T. Michigan, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explorations Newsletter*, vol. Volume 19, no. 2, pp. 25–35, December 2017.

[8] C. C. Aggarwal and P. S. Yu, Eds., *Privacy-preserving Data Mining: Models and Algorithms*. Springer, 2008.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] D. S. Carrell, D. J. Cronkite, M. R. Li, S. Nyemba, B. A. Malin, J. S. Aberdeen, and L. Hirschman, "The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1536–1544, 08 2019. [Online]. Available: https://doi.org/10.1093/jamia/ocz114

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[12] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965, pMID: 12261830. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775

[13] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," in *Proceeding of the workshop Privacy and Anonymity in the Information Society (PAIS) of the EDBT/ICDT 2016 Joint Conference*, T. Palpanas and K. Stefanidis, Eds., Bordeaux, France, March 15 2016.

[14] J. P. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. B. Cohen, and W. Duch, "A shared task involving multi-label classification of clinical free text," in *Biological, translational, and clinical language processing*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 97–104. [Online]. Available: https://www.aclweb.org/anthology/W07-1013

[15] K. Weilhammer, U. Reichel, and F. Schiel, "Multi-Tier Annotations in the Verbmobil Corpus," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May, 29–31, 2002.

[16] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. [Online]. Available: https://www.aclweb.org/anthology/H90-1021

[17] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *ArXiv*, vol. abs/1805.10190, 2018.

[18] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," in *NAACL-HLT*, 2019.

[19] X. Li, S. Panda, J. Liu, and J. Gao, "Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems," *arXiv preprint arXiv:1807.11125*, 2018.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.