

Towards Quantifying The Privacy Of Redacted Text

Vaibhav Gusain¹[0000–0002–7008–5201] and Douglas Leith¹[0000–0003–4056–4014] *

Trinity College Dublin, Ireland

Abstract. In this paper we propose use of a k-anonymity-like approach for evaluating the privacy of redacted text. Given a piece of redacted text we use a state of the art transformer-based deep learning network to reconstruct the original text. This generates multiple full texts that are consistent with the redacted text, i.e. which are grammatical, have the same non-redacted words etc, and represents each of these using an embedding vector that captures sentence similarity. In this way we can estimate the number, diversity and quality of full text consistent with the redacted text and so evaluate privacy.

Keywords: Transformers · Text privacy · Data leaks · k-anonymity

1 Introduction

Redacting a piece of text involves replacing selected words with an uninformative mask symbol. Redaction is widely used, but is generally carried out manually and there has been little analysis of the degree of privacy obtained. Note that evaluating text privacy is generally not straightforward since even when a word is redacted it might still be possible to reliably estimate it from the surrounding text i.e. the context of the redacted word may be revealing.

Machine learning models for text embedding are often trained by masking out individual words in a piece of text and selecting a model that best reconstructs the missing text. The idea here is that similar words appear in a similar context. In particular, transformer-based neural networks such as BART [6] adopt this approach and achieve state of the performance in many natural language processing tasks.

Given a piece of redacted text, in this paper we apply transformer-based neural networks to try to reconstruct the original text. For example, when the text he was stationed at singapore is redacted to he was stationed at <mask> then the top 5 reconstructed text predictions by BART are shown in Table 1. It can be seen that the reconstructed text is grammatical, consistent with the redacted text (has the same non-redacted words etc) and plausible even though in this example it does not correctly predict the missing word.

In this paper we study using such predicted reconstructions as the basis for a quantitative privacy metric for redacted text. This is motivated by the observation that the number of reconstructions that are estimated with high confidence

* This work was supported by Science Foundation Ireland grant 16/IA/4610.

		Redacted sentence	BART top prediction
he was stationed at <mask>		<mask> was	This article was originally published at singapore
he was stationed at the	0.62	<mask> at singapore	A look at singapore
he was stationed at:	0.58	<mask> <mask>	
he was stationed at Gettysburg	0.49	<mask> at singapore	
he was stationed at Ft.	0.48	<mask> <mask>	Singapore singapore
he was stationed at Knox	0.47	singapore	

Table 1: Left-hand table: Top 5 reconstructions by BART for the redacted sentence he was stationed at <mask>. The values shown in the second column are the corresponding confidence values output by BART. Right-hand table: top prediction by BART as the number of redacted words is increased.

can be expected to provide an approximate k-anonymity [9] measure i.e a measure of "Hiding in the crowd" privacy since there are at-least K sentences that are plausibly consistent with the redacted text. Since the reconstructions are represented as embedded vectors that capture sentence similarity (similar sentences are represented by nearby vectors) then we can also estimate the diversity of the reconstructions.

This work reported here is just a first, exploratory step but we find that this general approach shows promise.

Rather than evaluating k-anonymity and text diversity, we begin by considering the text quality of the predictions since this turns out to be a useful predictor of privacy in coarse classification tasks such as sentiment analysis, news article categorisation and medical condition (e.g. has cancer or not). We find that there is a thresholding effect, whereby beyond a certain level of redaction the quality tends to drop sharply. By carrying out simulated attacks against the redacted text we find that the drop in BART prediction quality strongly correlates with a decrease in attack effectiveness. The proposed approach therefore has the potential to provide a practical, useful estimate of redacted text privacy.

1.1 Related Work

Text Redaction. Despite the widespread use of redaction, there has been very little work on quantifying the privacy of redacted text or on evaluating robustness to attacks that seek to generate privacy leaks. Instead most work to date has focused on identifying personal data with text so that it can be redacted. See, for example, [2] which considers discovery of names, home towns etc in student discussion boards, and also the references therein. The closest work to the present paper is probably [1] which considers randomly redacting words to ensure a form of differential privacy and evaluates utility using a transformer neural net. However, there is no evaluation of the robustness of the redacted text to adversarial attacks (which is primarily what we use transformer neural nets for here) and

the interpretation of differential privacy in the context of redacted text remains unclear (in [1] the surrounding context of a redacted word is ignored, yet will often have an important impact on the degree of privacy achieved).

Text Reconstruction. Predicting missing text has been the subject of a great deal work in recent years. The state of the art uses transformer-based neural net architectures, following the breakthrough performance achieved by BERT. BART [6] is a transformer-based neural net that targets reconstruction of text damaged by spelling mistakes, missing words etc. Roughly speaking it is an amalgamation of BERT and GPT2, consisting of a bidirectional encoder which is very similar to BERT and a left-to-right decoder which is very similar to GPT2. This design allows BART to even predict arbitrary length of text for a single mask token which cant be achieved with BERT.

2 Quantifying BART Text Quality

The right-hand table in Table 1 shows how the top predicted sentence reconstruction by BART varies as the number of redacted words is increased. It can be seen that by the time four out the five words in the sentence are redacted the BART prediction degrades and is no longer grammatical. In our experiments (see below), we find that this behaviour is a common feature of the BART reconstructions. Of course, it is quite reasonable behaviour since at this point there is so little information left in the redacted sentence that BART has few clues as to how it might be reconstructed. Equally, the point where this information loss occurs is obviously also of great interest from a privacy viewpoint.

Rather than considering just the top prediction by BART, we proceed by considering the top N predictions, typically with N=100. We then estimate the fraction of these predictions which are not grammatical, and investigate the use of this as a measure of privacy.

In general, it is not trivial to estimate whether a sentence is grammatical or not. Fortunately we do not need to solve the general problem but can instead exploit the fact that BART predictions tend to either be fairly grammatical or else are grossly non-grammatical e.g. with many repetitions of the same word (as can be seen in Table 1) and/or with many repetitions of punctuation and spurious characters. That is, the BART predictions tend to either be reasonable text or to be “gibberish”.

To classify a sentence as gibberish or not, in our experiments we use Algorithm 1 although other choices are of course possible. Algorithm 1 combines a standard gibberish detector Nostrill [4] with a measure of the fraction of words from the original (non-redacted) sentence that overlap with the predicted sentence. Hyperparameter C controls the weight attached to each measure.

3 Experimental Measurements

3.1 Datasets Used

We evaluated performance on five datasets: four standard text classification datasets BBCnews [3], Amazon-Fine-food [8], AGnews [11], IMDB [7] plus the

Algorithm 1 Algorithm used to classify BART predictions as gibberish or not. Si is the actual input sentence without the mask, Sp is the BART prediction and C is a hyperparameter that checks the number of overlapping words between Si and Sp. It returns True if the prediction is estimated to be gibberish else it returns false

```

gibberish = use Nostrill to check if the Sp is gibberish or not.
if gibberish then
    return gibberish
else
    return customGibberish(Si,Sp,C)
end if

customGibberish(Si, Sp, C) :
Si  $\leftarrow$  number of uniques words in Si
Sp  $\leftarrow$  number of uniques words in Sp
p  $\leftarrow$  number of common words in Si and Sp / length(Si)
gibber = p*100/ length(Si)
if gibber  $\leq$  C then
    return True
else
    return False
end if
```

Medal medical dataset [10]. BBCnews has five classes (Business, Entertainment, Politics, Sport, Tech), Amazon-Fine-food has review stars and reviews with greater than 3 stars were assigned to one class and the rest to another class, AGnews has four classes (World, Sports, Business, Sci/Tech), IMDB has two classes (positive and negative sentiment), Medal has two classes (text specifically about cancer diseases, plus the rest). Each dataset was split 80:20 into a training and a test dataset, with the training dataset being available to the adversary but not the test dataset. The datasets are sampled so that they are balanced by category.

3.2 Threat Model

The attacker can observe redacted text, and a training data subset of each dataset. The redacted text is derived from held out data not available to the attacker. The aim of the attacker is to discover the category of the text e.g. for a movie review to discover the sentiment, for a news article to discover the news category.

3.3 Reconstruction Quality Metric

For each redacted sentence we take the top 100 reconstruction predictions from BART and apply Algorithm 1 to classify them as either gibberish or not, assigning value +1 for gibberish and 0 otherwise. We calculate the mean of these 100 values.

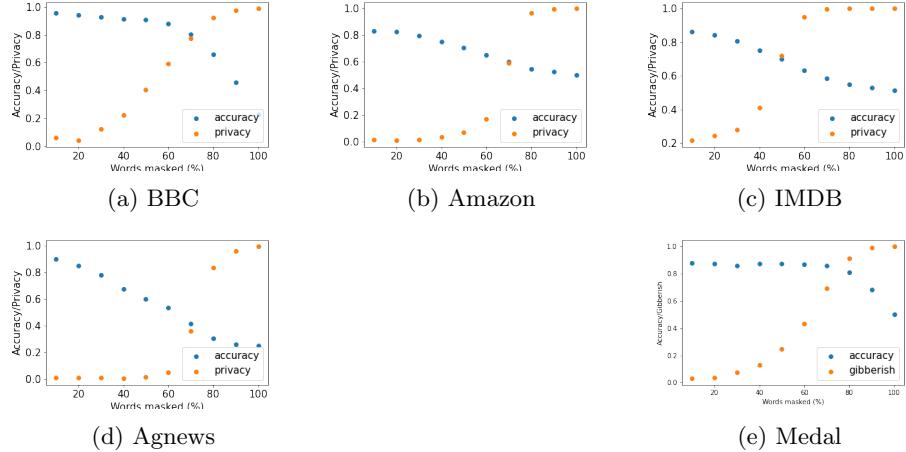


Fig. 1: Measured privacy metric and attack accuracy for each dataset as the fraction of redacted text is varied from 0 to 100%.

3.4 Privacy Attack Performance Metric

Using the training data for each dataset the adversary trains a classifier based on a TFIDF [5] vectoriser and a logistic regression model (for these datasets it is known that classifiers of this sort are able to achieve high accuracy). Given redacted text, the attacker then uses this classifier to estimate the category of the text. We evaluate the success of this reconstruction using the mean accuracy of these predictions i.e. the fraction of redacted sentences for which the category is correctly estimated. The test data is balanced, so accuracy is an informative performance measure.

3.5 Redaction Strategy

For each dataset we encode the words using a TFIDF vectoriser (discarding words with document frequency less than 10%). We then vary the level of redaction by replacing a random X percent of words by a mask token, varying X from 0 to 100%. Using TFIDF in this way avoid ineffectual masking of stop words and other uninformative words. Other redaction strategies are, of course, possible.

3.6 Additional Material

We will post our implementations and the associated data on github.

3.7 Results

Figure-1 shows the measured privacy metric and attack accuracy for each dataset as the fraction of redacted text is varied. It can be seen that as the percentage of

masked words is increased the classification accuracy decreases while the privacy metric increases.

When less than around 20% of words are redacted, the privacy metric is close to zero for every dataset, indicating that BART consistently reconstructs grammatical sentences that are consistent with the redacted text. Analysis of the top 100 BART predictions (not included here) show little diversity in the sense that the sentence embedding vectors tend to cluster together. The attack accuracy is correspondingly also consistently high.

When greater than around 80% of words are redacted, then the privacy metric is close to 100% and the attack accuracy is approximately the reciprocal of the number of categories i.e. comparable with a random coin toss.

Between 20 and 80% redaction the privacy metric increases and the attack accuracy correspondingly decreases. By selecting a level of redaction that ensures the privacy metric is above a target threshold, e.g. 70%, then these measurements indicate that a good level of robustness against the reconstruction attack can be obtained across a wide range of datasets.

3.8 Discussion

Due to lack of space we do not include an evaluation of utility here, which can be expected to degrade as privacy increases. However, we note briefly that we have evaluated next word prediction performance for the Medal dataset vs privacy and find that the utility remains high even when redaction achieves a high level of resistance against estimation of medical condition.

We use attack accuracy as a proxy for privacy, since it is difficult to apply standard privacy metrics such as k-anonymity and differential privacy to natural language text data. However, initial results indicate that it may be possible to estimate a metric similar to k-anonymity by clustering the embedding vectors of the BART predictions and counting the number of distinct clusters. In the regime where BART predictions are grammatical (redaction level less than 20% in Figure-1) these clusters reflect semantic diversity, whereas in the regime where BART predictions produce lower quality text the clusters tends to become less informative. However, we leave proper analysis of these aspects to future work.

Initial results also suggest that the nature of the privacy threat is relevant to the level of redaction needed. To prevent disclosure of broad textual aspects such as sentiment or new category our results show that a high level of redaction is necessary, but preventing disclosure of more fine-grained aspects might be achievable with lower levels of redaction. Again, we leave further study of this to future work.

References

1. Adelani, D.I., Davody, A., Kleinbauer, T., Klakow, D.: Privacy guarantees for de-identifying text transformations. arXiv preprint arXiv:2008.03101 (2020)
2. Bosch, N., Crues, R., Shaik, N., Paquette, L.: "hello,[redacted]": Protecting student privacy in analyses of online discussion forums. Grantee Submission (2020)

3. Greene, D., Cunningham, P.: Practical solutions to the problem of diagonal dominance in kernel document clustering. In: Proc. 23rd International Conference on Machine learning (ICML'06). pp. 377–384. ACM Press (2006)
4. Hucka, M.: Nostril: A nonsense string evaluator written in python. Journal of Open Source Software **3**(25), 596 (2018). <https://doi.org/10.21105/joss.00596>, <https://doi.org/10.21105/joss.00596>
5. Jing, L.P., Huang, H.K., Shi, H.B.: Improved feature selection approach tfidf in text mining. In: Proceedings. International Conference on Machine Learning and Cybernetics. vol. 2, pp. 944–946 vol.2 (2002). <https://doi.org/10.1109/ICMLC.2002.1174522>
6. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
7. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), <http://www.aclweb.org/anthology/P11-1015>
8. McAuley, J.J., Leskovec, J.: From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In: Proceedings of the 22nd international conference on World Wide Web. pp. 897–908 (2013)
9. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1998)
10. Wen, Z., Lu, X.H., Reddy, S.: MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.clinicalnlp-1.15>, <https://doi.org/10.18653%2Fv1%2F2020.clinicalnlp-1.15>
11. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. CoRR **abs/1509.01626** (2015), <http://arxiv.org/abs/1509.01626>