Interpretable Privacy Preservation of Text Representations Using Vector Steganography

Geetanjali Bihani

Purdue University
gbihani@purdue.edu

Abstract

Contextual word representations generated by language models (*LMs*) learn spurious associations present in the training corpora. Recent findings reveal that adversaries can exploit these associations to reverse-engineer the private attributes of entities mentioned within the corpora. These findings have led to efforts towards minimizing the privacy risks of language models. However, existing approaches lack interpretability, compromise on data utility and fail to provide privacy guarantees. Thus, the goal of my doctoral research is to develop interpretable approaches towards privacy preservation of text representations that retain data utility while guaranteeing privacy. To this end, I aim to study and develop methods to incorporate steganographic modifications within the vector geometry to obfuscate underlying spurious associations and preserve the distributional semantic properties learnt during training.

Introduction

General purpose language models trained on large corpora of natural language text are prone to learning generalizations from the underlying data (Coavoux, Narayan, and Cohen 2018). Recent studies have shown that language variety contains cues that reveal private attributes of individuals, such as identity, political inclination and gender (Nguyen, Rosseel, and Grieve 2021). These cues allow adversaries to exploit the language models and uncover private attributes of users who speak, write or are mentioned within the text used to train the models (Coavoux, Narayan, and Cohen 2018). These findings have led to an increased interest in privacy preservation research for natural language processing.

Research Problem

Current privacy preservation methods for NLP rely on the irreversible removal of rich social signals from data, rendering a normalized text representation geometry (Nguyen, Rosseel, and Grieve 2021). These methods can be divided into three groups: (1) Differential privacy based models (2) Cryptographic models and (3) Adversarial Learning models. While adversarial privacy preservation fails to address explicit leakages, differential privacy preservation fails to defend against implicit leakages (Huang et al. 2020). In theory,

provable privacy defense against both implicit and explicit leakage can be provided by cryptographic methods such as multi-party computations (*MPC*) and homomorphic encryption, but due to their computational burden and loss of data utility, these methods have been been deemed inefficient to be applicable within the deep learning paradigm (Huang et al. 2020). To combine the benefits of cryptography and deep learning, Huang et al. (2020) recently proposed adding an instance encoding step while fine-tuning language models in a federated learning setting. Although this approach provided better data utility and faster computations than *MPC* and homomorphic encryption (Huang et al. 2020), it has been shown to succumb to reconstruction attacks (Carlini et al. 2021). These findings suggest that although cryptographic enhancements to deep learning models might provide significant benefits in terms of privacy guarantees, their current adaptations lack reliability, data utility and computational efficiency. Furthermore, while partially meeting privacy objectives, these models lack explainability and interpretability, failing to concurrently meet fairness and trust objectives within ethical NLP (Pruksachatkun et al. 2021).

In the past, privacy preservation in various media has been achieved through information hiding, using various cryptographic techniques (Krishnan, Thandra, and Baba 2017). Steganography is one such technique, which hides secret information within a given medium, without altering the structure of the medium itself. To our knowledge, text steganography approaches for privacy preservation have been limited to text character or bit manipulation, because minor modifications in the text form lead to semantic anomalies (Krishnan, Thandra, and Baba 2017). For multi dimensional representations, steganalysis methods such as one described in Jiang (2008) use intensity and imperceptibility features to cover and uncover secret messages within the vector space. Since language models map individual word occurrences to multi-dimensional text representations, utilizing steganography in multi-dimensional spaces for privacy preservation in text representations still remains unexplored.

Completed Work

Our prior work on analyzing text representation geometries showed that developing a general criteria to target spurious associations in text representations is a challenge, largely due to the dependence of distributional text representation

spaces on the underlying model choices and training corpora (Bihani and Rayz 2020, 2021). Our study of attributive word associations in text representation models revealed that models trained on the same text corpora using different context aggregation schemes encode different word associations (Bihani and Rayz 2020). In another recent paper (Bihani and Rayz 2021), we found that general purpose language models encode varying degrees of word-sense associations across model layers. Specifically, bidirectional language models encode stronger associations between the same polysemes of word occurrences as compared to unidirectional models. Moreover these associations are better retained in lower model layers and are significantly reduced in the upper layers due to contextualization of representations. These results indicate that we cannot create an exhaustive list of the word associations that need to be targeted for privacy preservation *a priori*.

Plan of Research

To address the gaps in prior works and design trustworthy models of privacy preservation in NLP, I aim to utilize steganography within the text representation space to concurrently preserve data utility and privacy. Hence, I propose the following research question for my doctoral research; ***RQ:*** *How effectively can steganographic models of privacy preservation obfuscate spurious associations (implicit and explicit) without compromising data utility?* My goal is to minimize the normalization of signals learnt from the training data while ensuring privacy on demand. I propose to achieve this by combining steganography and micro-aggregation within the vector space to obfuscate spurious associations in the vector space geometry i.e. Vector space steganography (*VSS*). This approach will allow the retention of the distributional semantics signals learnt from the data, maintaining data utility and accuracy, while hiding the latent confounds within the vector space, allowing privacy preservation. Moreover, *VSS* will allow interpretability, working towards building trustworthy privacy preserving NLP techniques.

Towards this goal, I define three research objectives, which will be completed according to the tentative timeline in Table 1. My first research objective is–**RO1:** *Designing VSS applicable on-top-of general purpose language model representations to obfuscate spurious associations.* This objective addresses the privacy preservation part of our research question. To obfuscate any spurious associations related to a given word *w*, steganographic candidates will be generated from the vector space geometry by reversing the steganalysis criteria described for multidimensional vectors in (Jiang 2008). These candidates will be micro-aggregated to create an obfuscated representation for *w*, hiding any latent confounds associated with *w*. My second research objective is–**RO2:** *Evaluating the data utility post-VSS across general purpose language model representations.* This objective addresses the data utility aspect of my research question. To measure data utility, I will compare the proposed approach to the baseline accuracy thresholds for various NLU tasks as described in (Huang et al. 2020). My final research objective is–**RO3:** *Evaluating the provable privacy guaran-*

Research Objective	Time Period
RO1	August 2021 - November 2021
RO2	December 2021 - March 2021
RO3	April 2022 - July 2022

Table 1: Timeline for proposed research plan

tees of VSS. This objective will address the efficiency of *VSS* in terms of its privacy guarantees, where we will consider the definitions and baselines provided in (Carlini et al. 2021). Through these objectives, I aim to make progress towards developing methods of interpretable privacy assurance for text representation geometries.

References

Bihani, G.; and Rayz, J. 2021. Low Anisotropy Sense Retrofitting (LASeR) : Towards Isotropic and Sense Enriched Representations. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 81–95. Online: Association for Computational Linguistics.

Bihani, G.; and Rayz, J. T. 2020. Model Choices Influence Attributive Word Associations: A Semi-supervised Analysis of Static Word Embeddings. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 568–573.

Carlini, N.; Deng, S.; Garg, S.; Jha, S.; Mahloujifar, S.; Mahmoody, M.; Thakurta, A.; and Tramèr, F. 2021. Is Private Learning Possible with Instance Encoding? In *2021 IEEE Symposium on Security and Privacy (SP)*, 410–427. IEEE.

Coavoux, M.; Narayan, S.; and Cohen, S. B. 2018. Privacy-preserving Neural Representations of Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1–10.

Huang, Y.; Song, Z.; Chen, D.; Li, K.; and Arora, S. 2020. TextHide: Tackling Data Privacy in Language Understanding Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1368–1382. Online: Association for Computational Linguistics.

Jiang, N. 2008. A Novel Analysis Method of Information Hiding. In *2008 Congress on Image and Signal Processing*, volume 5, 621–625.

Krishnan, R. B.; Thandra, P. K.; and Baba, M. S. 2017. An overview of text steganography. In *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, 1–6.

Nguyen, D.; Rosseel, L.; and Grieve, J. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 603–612. Online: Association for Computational Linguistics.

Pruksachatkun, Y.; Ramakrishna, A.; Chang, K.-W.; Krishna, S.; Dhamala, J.; Guha, T.; and Ren, X. 2021. Proceedings of the First Workshop on Trustworthy Natural Language Processing. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*.