Privacy-Preserving application using Homomorphic Encryption on NLP algorithms

Jose Contreras
University of Tartu
jose.angel.contreras.gedler@ut.ee

Karlos Taaniel Lillemäe
University of Tartu
karlos.taaniel.lillemae@ut.ee

1 Introduction

Today's increasing use of Artificial Intelligence (AI) and Natural Language Processing (NLP) tools makes data privacy relevant, given that these tools often handle sensitive information. Therefore, it is necessary to include the secure handling of data as part of the design of any new AI tool, bearing in mind the existence of malicious entities capable of detecting system vulnerabilities and infiltrating information to carry out their purposes.

Moreover, for an AI project to gain the trust of its users, it must commit to safeguarding the privacy of the data shared by them. Scientific research in this field is very fertile, given the large amount of data circulating on the web every minute, much of which represents sensitive information that must be protected.

It is necessary to analyze the data processing stages to determine which privacy preservation techniques are best suited to each system, always seeking to ensure that the information provided by users is treated securely in all steps of the design and development of any system.

2 Related Work

A proposal that seeks to improve data privacy and protection with a methodology that uses NLP and unsupervised machine learning was made by (Martinelli et al., 2020) to adequately manage sensitive and personal information.

The objective of (Martinelli et al., 2020) was to create a valuable set of labeled data for training supervised machine learning algorithms to detect and classify sensitive privacy information in texts. The authors worked on Italian data and used a total of 1000 documents for the study, belonging to the health and justice sectors, 500 documents for each case, representing a minimal amount of data. However, the results were encouraging, and the authors suggest continuing in this direction to obtain new advances.

(Klymenko et al., 2022) address the alternative of using differential privacy in NLP. The authors clarify that this option is promising but has tremendous challenges. First, the data to be worked on must be evaluated or selected. The data will be unstructured, and it must be precisely known which information should be protected to avoid unnecessary computational expenses in protecting trivial data.

The mathematical solidity on which differential privacy is built has made it a good option for addressing security in this area. Additionally, it works on structured databases that contain the information that needs to be protected. (Klymenko et al., 2022) also present a generalization of differential privacy, known as metric differential privacy, which is a better alternative for NLP. They explain that when applying differential privacy, there is a perturbation of the original data, which designers control to work within a threshold that, despite modifying the original data to provide privacy, allows for practical work to obtain relevant information related to the original dataset, without compromising the privacy of the data.

An interesting approach was taken by (Aono et al., 2016) by combining logistic regression with homomorphic encryption. Logistic regression is efficient in machine learning for classifying data, and the task of protecting such data using homomorphic encryption guarantees the integrity and confidentiality of the information.

Homomorphic Encryption (HE) is a cryptographic technique that allows the processing of encrypted data without decrypting it (Armknecht et al., 2015). This technique is based on the fact that the result of some mathematical operations on encrypted data is the same as if the operation were performed on the original data. This property allows mathematical operations on encrypted data without decrypting it.

The contribution of (Aono et al., 2016) is the design of a secure and scalable system. They worked on the original logistic regression, transforming it into an equivalent regression using homomorphic encryption. Additionally, they have added instructions to add differential privacy to the system.

Recognizing the existence of reverse engineering tools that can extract the original information from systems designed with embedding techniques, (Lee et al., 2022) proposed using homomorphic encryption on original data. Pre-trained BERT embeddings were used in their article, allowing the training of a logistic regression classifier. The results indicate that working with encrypted data enhances privacy and can assure users that malicious entities will use no private information. To this end, local privacy configuration was used, in which the user encodes their information before submitting it to a specific service provider.

Web page phishing remains a problem for cybersecurity as attackers continually refine their techniques, and users remain vulnerable to scams. (Chou et al., 2020) analyzed images corresponding to both fake and legitimate web pages to establish their visual similarities and differences in order to differentiate them. In terms of privacy, they designed a system that uses homomorphic encryption and cloud computing, allowing the user to interact with the system by sending an encrypted image of the web page they wish to access to the cloud, which is examined to determine whether it is malicious or not. If found malicious, the user will receive an alert, avoiding any information theft or being scammed.

What is relevant for us from (Chou et al., 2020) is the threat model, where the parties involved in the inference are semi-honest actors, which means that the attacker has access to the data but cannot modify it. (Feng et al., 2021) follow this approach combining HE and Garbled Circuits into a hybrid structure for inference in Recurrent Neural Networks (RNN), a model called CryptoGRU. The model was tested on public datasets like Enron and IMDB, showing that the model can achieve high accuracy and latency.

THE-X is a tool that uses homomorphic encryption on transformers (Chen et al., 2022). It is a practical application that allowed them to conclude that while this type of encryption risks a percentage of performance, the privacy of the data is theoretically proven. Good results were obtained in sentiment analysis, paraphrasing, and text classification, and in future work, they plan to address performance issues.

There are many challenges in using homomorphic encryption in NLP, and (Chen et al., 2022) summarizes the replacements needed to practically incorporate HE into modern NLP models. In particular, the most relevant issue is that HE operates only with basic additions and multiplication over the integers, or more precisely, over the ring of integers modulo a prime number. Thus, the model must be modified to work with integers, and the weights and biases must be quantized to integers. Despite the costs that this model imposes in terms of performance, this is one of the first practical applications of HE in transformers inference.

References

Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. 2016. Scalable and secure logistic regression via homomorphic encryption. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, CODASPY '16, pages 142–144, New York, NY, USA. Association for Computing Machinery.

Frederik Armknecht, Colin Boyd, Christopher Carr, Kristian Gjøsteen, Angela Jäschke, Christian A. Reuter, and Martin Strand. 2015. A guide to fully homomorphic encryption. Cryptology ePrint Archive, Paper 2015/1192. https://eprint.iacr.org/2015/1192.

Tianyu Chen, Hangbo Bao, Shaohan Huang, Li Dong, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. 2022. THE-X: Privacy-preserving transformer inference with homomorphic encryption. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3510–3520, Dublin, Ireland. Association for Computational Linguistics.

Edward Chou, Arunkumar Gururajan, Kim Laine, Nitin Kumar Goel, Anna S. Bertiger, and Jack W. Stokes. 2020. Privacy-preserving phishing web page classification via fully homomorphic encryption. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2792–2796.

Bo Feng, Qian Lou, Lei Jiang, and Geoffrey Fox. 2021. CRYPTOGRU: Low latency privacy-preserving text analysis with GRU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2052–2057, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*. Association for Computational Linguistics.

Garam Lee, Minsoo Kim, Jai Hyun Park, Seung won Hwang, and Jung Hee Cheon. 2022. Privacy-preserving text classification on bert embeddings with homomorphic encryption.

Fabio Martinelli, Fiammetta Marulli, Francesco Mercaldo, Stefano Marrone, and Antonella Santone. 2020. Enhanced privacy and data protection using natural language processing and artificial intelligence. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.