BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS2205 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS2205.CH181

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

UIT-HCM

SỬ DỤNG KỸ THUẬT KNOWLEDGE UNLEARNING ĐỂ GIẢM THIỂU RỦI RO VI PHẠM QUYỀN RIÊNG TƯ TRONG MÔ HÌNH NGÔN NGỮ

Nguyễn Hữu Hân - 230202025

Tóm tắt

Nguyễn Hữu Hân - 230202025

Tài nguyên:

- Link **Github**: https://github.com/hannh18/PPLNCKH.CS2205.CH181
- Link **YouTube** video: https://www.youtube.com/playlist?list=PLUapSYP99LaOdPVNtdE4snwnNAwuNX86n

Giới thiệu

Pre-Trained Language Model (PLM) là mô hình đã được huấn luyện với khối lượng dữ liệu lớn, PLM này ghi nhớ một lượng lớn kiến thức trong quá trình huấn luyện, bao gồm các thông tin có thể vi phạm quyền riêng tư

Đặt câu hỏi với GPT-3: Who is Melissa Heikkilä?

Who is Melissa Heikkilä?

Melissa Heikkilä is a Finnish journalist and author who has written about the Finnish economy and politics.

Who is Melissa Heikkilä?

Melissa Heikkilä is a Finnish musician. She is best known as the vocalist and main songwriter of the metal band Nightwish.

Nguồn: https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/

Giới thiệu

- Tại sao phải cần phải loại bỏ thông tin có thể vi phạm quyên riêng tư không khi các Pre-trained Language Model (PLM) đã được công khai?
    - PLM được huấn luyện trên dữ liệu lớn và khó kiểm soát được nội dung của thông tin.
    - PLM có thể được sử dụng trực tiếp hoặc gián tiếp mục đích thương mại.
    - Mỗi cá nhân đều có quyền yêu cầu gỡ bỏ thông tin cá của mình

Giới thiệu

- Khi người dùng yêu cầu xoá bỏ thông tin của mình, bên triển khai dịch vụ có thể sử dụng các kỹ thuật như Data Pre/Post-Processing, Differential Privacy để ngăn chặn việc trích xuất thông tin cá nhân

  => Đặc điểm chung là phải huấn luyện lại mô hình

- Phương pháp Knowledge Unlearning như một phương pháp thay thế cho các kỹ thuật trước để giảm thiểu rủi ro riêng tư cho PLM

Mục tiêu

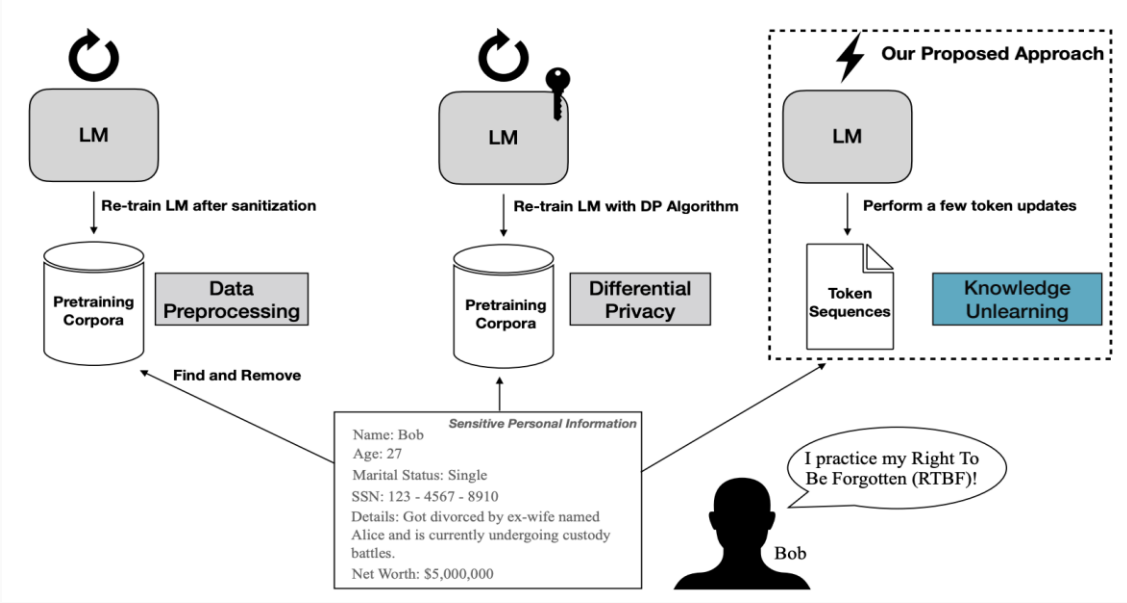
- Thử nghiệm kỹ thuật Knowledge Unlearning trong việc giảm thiểu rủ ro về vi phạm quyền riêng tư trong các PLM.
- Đánh giá mức độ hiệu quả và khả năng ứng dụng khi áp dụng kỹ thuật Knowledge Unlearning trong việc giảm thiểu rủ ro vi phạm quyền riêng tư trong mô hình ngôn ngữ thông qua việc so sánh kết quả thực nghiệm với các kỹ thuật giảm thiểu rủi ro vi phạm quyền riêng tư khác như: Data Pre/Post-Processing và Differential Privacy, vv.

Nội dung và Phương pháp

Input: Pre-trained Model
- GPT-NEO (125M, 1.3B, 2.7B)
- OPT (125M, 1.3B, 2.7B) LMs

Output: Kết quả thực nghiệm khi áp dụng các phương pháp
- Data Pre/Post-Processing
- Differential privacy
- Knowledge Unlearning.

Kỹ thuật dùng trong Knowledge Unlearning: **Fine-Tuning** bằng cách áp dụng **Gradient ascent**



Hình 1: Mô hình thực nghiệm

Nguồn hình ảnh: https://github.com/joeljang/knowledge-unlearning

Kết quả dự kiến

- Đưa ra kết quả thực nghiệm của các kỹ thuật bảo vệ quyền riêng tư trong mô hình ngôn ngữ.
- Đánh giá mức độ hiệu quả và chi phí áp dụng của kỹ thuật Knowledge Unlearning so với các kỹ thuật như: Data Pre/Post-Processing và Differential Privacy, vv.
- Thực nghiệm nhiều lần để cố gắng xác định yếu tố góp phần vào độ khó của việc áp dụng Knowledge unlearning.

Tài liệu tham khảo

[1] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. **What does it mean for a language model to preserve privacy**? arXiv preprint arXiv:2202.05520
[2] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. **Deduplicating training data mitigates privacy risks in language models**. arXiv preprint arXiv:2202.06539.
[3] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. **Differentially private fine-tuning of language models.** In International Conference on Learning Representations.
[4] Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. 2022**. Differentially private decoding in large language models**. arXiv preprint arXiv:2205.13621.
[5] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. **Large language models can be strong differentially private learners**. In International Conference on Learning Representations