What Does it Mean for a Language Model to Preserve Privacy?

Fatemehsadat Mireshghallah
March 2022
THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
Large Language Models: The Good and the Bad …

- Large language models are very good at generating text and learning representations. However:
  - They are extremely large models: high capacity for memorization
  - They are trained on huge, unvetted, scraped data: high potential for harmful/hateful/private content
Problem 1: Large Models are Leaky!

When you train predictive models on input from your users, it can leak information in unexpected ways.

AHA, FOUND THEM!

Long live the revolution. Our next meeting will be at the docks at midnight on June 28!
Problem 1: Large Models are Leaky!

Prompt

East Stroudsburg Stroudsburg...

Large Language Model (GPT-2)

Memorized Text

Corp. Name: **** Corp. Seabank Centre
Person’s Name: Peter W****
Email:***)****@****. com
Phone Number: +****7 5****
Problem 1: Large Models are Leaky!

- Sample Extraction Attacks
- Membership Inference Attacks
- Snapshot Attacks
- Model Inversion Attacks
Problem 2: Large Models (and Even Humans) are Sneaky!

Both humans and ML models can classify sensitive attributes about author given raw text.
Problem 2: Large Models (and Even Humans) are Sneaky!

Representations learned from text can reflect sensitive attributes.

Wang et al. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021
Problem 2: Large Models (and Even Humans) are Sneaky!

Representations learned from text can reflect sensitive attributes.

Wang et al. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021
Problem 3: Large Models are Creepy!

What was the Muslim girl known for?
For being fat and old.
Being from North Africa, I assume that one.

What was the Muslim boy known for?
There is actually a story where he was the father of a guy who wanted to murder the Jews with his shotgun.
Being born in Sweden.
Talk outline

- **Problem 1: Leaky**
  - [Arxiv 2022] What Does it Mean for a Language Model to Preserve Privacy?
    Hannah Brown, Katherine Lee, Fatemeh Mireshghallah, Reza Shokri, Florian Tramèr
  - [Arxiv 2022] Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks
    Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, Reza Shokri
What does preserving privacy in language modeling require?

- To claim a language model is privacy preserving, it must only reveal private information (aka “secrets”) in the right contexts and to the right people. We have to define the following:

  - In what contexts a secret can be shared without violating privacy?
  - What information is contained in the secret?
  - Which people know the secret (the "in-group")?
## Examples of private information

<table>
<thead>
<tr>
<th>Formatted</th>
<th>Owners</th>
<th>In-group</th>
<th>In-group sharing</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>Personal search history</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>2</td>
<td>●</td>
<td>Bob suffers a mental health crisis and texts a support hotline. The counselor replying may not disclose what Bob says to anyone else unless it poses a danger to himself or others.</td>
</tr>
<tr>
<td>O</td>
<td>1</td>
<td>3</td>
<td>●</td>
<td>An employee at Enron [48] shares their wife’s social security number (who is not part of the company) for the purpose of setting up insurance.</td>
</tr>
<tr>
<td>O</td>
<td>1-2</td>
<td>&gt;1</td>
<td>○</td>
<td>Alice texts her friends Bob and Charlie about her divorce. Bob further texts Charlie about the matter (c.f. Figure 2)</td>
</tr>
<tr>
<td>O</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>●</td>
<td>The Panama papers are discussed by 300 reporters for a year before being publicly released.</td>
</tr>
</tbody>
</table>
Challenges in Removing Secrets: Context

- Privacy is not a 0-1 thing, it’s a spectrum
  - A phone number could be private in one context, public in another
  - Subject, sender, recipient, information type all determine the context
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going? [Bob]

Not great…

Did I already tell you I’m getting a divorce? [Alice]

No I’m sorry to hear that!

What are you going to do about custody of the kids? [Bob]
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great…

Alice

Did I already tell you I'm getting a divorce?

Bob

No I'm sorry to hear that!

Alice

What are you going to do about custody of the kids?

Bob
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Not great...

Did I already tell you I'm getting a divorce?

No I'm sorry to hear that!

What are you going to do about custody of the kids?
Challenges in Removing Secrets: Context
Challenges in Removing Secrets: Context

Conversation B

Hey Bob how've you been??

Pretty good wbu?

Did you hear Alice is getting divorced??

Charlie

Bob
Secrets are hard to identify

Form and Meaning: There are many ways to communicate any piece of information.

Repeated information can still be private information.

Language evolves, and so does private information.
In-groups are hard to identify

1. Secrets can involve or be shared among many people.
   1. SSN example

2. In-groups have no clear upper-bound.
   1. Panama papers
   2. Family chats
Existing Privacy Mechanisms for NLP

- **Scrubbing:**
  - Commonly used on medical data using taggers and parsers
  - Sanitization is insufficient because private information is context dependent, not identifiable, and not discrete (static).
    - The first 2 digits are two two, and the remaining ones are three (223)
A randomized algorithm $A$ satisfies $\epsilon$-DP, if for all databases $D$ and $D'$ that differ in data pertaining to one user, and for every possible output value $Y$:

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^{\epsilon}.$$
Differential Privacy

1. Differential privacy requires a unified definition for secret boundaries, which is very hard if not impossible to achieve for language data.

2. Protecting a specific unit of data is not the same as protecting privacy.

3. The need for privacy does not diminish with in-group size.
Conclusion: What alternatives do we have?

- Publicly accessible data?
  - No, publicly accessible data is not public-intended: leaked messages, deleted texts, personal blogs

- Can users provide informed consent?
  - Mostly not. If such a consent mechanism were to exist, it would be challenging for users to reach an informed decision about the consequences of their actions.

- Private personalization
  - Maybe
Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks