Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M20-161P^
Published: 2/18/2021

Text Representation Learning for Preserving Textual Privacy and Utility of User Data

Inventors

Ghazaleh Beigi

Kai Shu

Ruocheng Guo

Suhang Wang

Huan Liu

Contact

Shen Yan
shen.yan@skysonginnovations.com

Background

Textual information is one of the most significant portions of data that users generate by participating in different online activities such as leaving online reviews and posting tweets. On one hand, textual data consists of abundant information about users' behavior, preferences, and needs which is critical for understanding them. For example, textual data has been used by service providers to track users' responses to products and provide them with personalized services. On the other hand, publishing intact user-generated textual data places user privacy at risk. The textual data itself contains sufficient information to re-identify users in the textual database and jeopardize their private attribute information.

In response to these privacy concerns, data publishers seek to protect users' privacy by anonymizing the data before sharing it. However, traditional privacy preserving techniques such as k-anonymity and differential privacy are inefficient for user-generated textual data because this data is highly unstructured, noisy, and unlike traditional documental content, consists of numerous short and informal prose. These techniques may impose a significant utility loss for protecting textual data as they may not explicitly include utility into their design objectives.

Invention Description

Researchers at Arizona State University have developed a novel double privacy-preserving text representation learning framework, DPText, which learns a textual representation that (1) is differentially private, (2) does not contain private information and (3) retains high utility for the given task. DPText consists of an auto-encoder for extracting text latent representation, a differential-privacy-based noise adder, and two discriminators for preserving semantic meaning and privacy. In two natural language processing tasks—sentiment analysis and part of speech tagging—DPText demonstrated effective preservation of both privacy and utility.

Potential Applications

- Social media
- Cybersecurity

- Textual data processing

Benefits and Advantages

- Protects users against re-identification and private-attribute inference attacks
- Preserves semantic meaning of user text
- Balances privacy and utility of user data

Faculty Homepage of Professor Huan Liu