TPRPF: A Preserving Framework of Privacy Relations Based on Adversarial Training for Texts in Big Data

Yuhan CHAI, Zhe SUN, Jing QIU,
Lihua YIN, Zhihong TIAN

Frontiers of Computer Science, DOI: 10.1007/s11704-022-1653-0

Problems & Ideas

- Problems:
  - The privacy relationship is related to specific other people, and it is the privacy that both users need to protect together. The disclosure of privacy relationships damages the interests of individuals and damages the interests of others; privacy relations are more needed and harder to protect.
  - Existing approaches protects only privacy attributes information about user, and doesn't consider the privacy relations between users contained in text representations.
- Ideas: The adversarial training of multi-party games is used to enhances the privacy of text representation and prevents attackers from inferring the privacy relations between users. Meanwhile, it is less affect the performance of primary learning tasks.

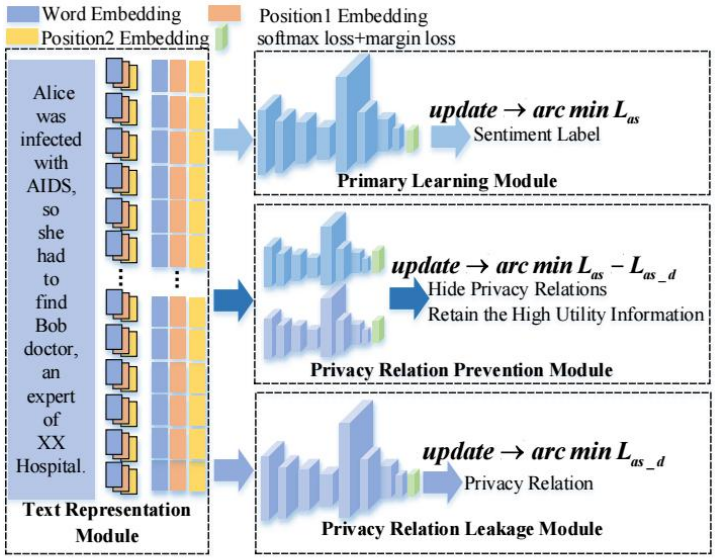

Fig. 2 The architecture of TPRPF framework.

Main Contributions

- Contributions:
  - We presented a Preserving Framework of Privacy Relations based on adversarial training for Texts in Big Data (TPRPF). It enhances the privacy of text representation and prevents attackers from inferring the privacy relations between users.
  - We applied the margin loss function to the softmax loss function and investigated the applicability and usefulness of the ConvMS as a classifier to better learn distinguishable deep features that improve the classification accuracy of the primary learning task;
  - We treat relations between users as privacy tasks that need to be protected by treating sentiment classification as an example of the primary learning task. We empirically show the effectiveness of the TPRPF framework and ConvMS model.

Table 1 Experimental Results of ConvMS Model

Model	s	m	w	*PL Accuracy(%)*	v	*RE Accuracy(%)*
ConvMS^{Softmax}	-	-	-	95.341	-	63.435
ConvMS	10	0.003	0.9	96.293	0.9	64.985

Table 2 Experimental Results of TPRPF Framework

Method	Privacy Parameters	Disturbance Thresholds	*PL* Accuracy(%)	*RE* Accuracy(%)
TPRPF^{Without}	-	-	96.293	64.226
TPRPF^{Adv}	-	0.1	95.422	62.952
TPRPF	1.0	-	96.763	51.059