The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

An automated privacy information detection approach for protecting individual online social network users

Weihua Li[*1], Jiaqi Wu[*1] Quan Bai[*2]

[*1] Auckland University of Technology, New Zealand [*2] University of Tasmania, Australia

Abstract: Massive private messages are posted by online social network users unconsciously every day, some users may face undesirable consequences. Thus, many studies have been dedicated to privacy leakage analysis. Whereas, there are very few studies detect privacy revealing for individual users. With this motivation, this paper aims to propose an automated privacy information detection approach to effectively detect and prevent privacy leakage for individual users. Based on the experimental results and case studies, the proposed model carries out a considerable performance.

1. Introduction

Online social networks (OSNs) have become ubiquitous in people's activities. The popularization of OSNs turns out to be a double-edged sword. On one hand, it provides convenience for people to communicate, collaborate, and share information. On the other hand, OSNs also come with serious privacy issues. Without given much attention by the users, a massive amount of private information can be accessed publicly through OSNs. Users may expose themselves to a wide range of "observers", which include not only relatives and close friends, but also strangers and even stalkers. This raises a serious cybersecurity issue, i.e., online privacy leak.

Online privacy leak means that an individual user shares his/her private information to people who he/she does not know well or even strangers on the Internet. This can be very dangerous for general Internet users, especially with the booming of OSNs. It is necessary to have a tool to assist general users to make better use of OSNs and protect them from leaking privacy information [Wang 11] [Hasan 13]. Hence, it is essential to detect privacy leakage in OSNs and remind individual online social network users before posting any privacy-related message. Under this motivation, in this paper, we propose a novel privacy detection framework for individual users of OSNs by using a Deep Learning approach. Twitter has been used as the source of data for training and validating our proposed framework since it is the biggest microblogging social media in the world [Mao 11]. Based on the generic definition of privacy and the characteristics of OSNs, the definition of "individual privacy" in OSNs have been formally defined. Furthermore, a deep learning-based approach has been developed and utilized to extract privacy-related entities from the messages posted by the users.

The rest of the paper is organized as follows. Section 2 reviews the existing research work regarding data leaks on OSNs. Section 3 introduces the automated privacy information detection framework. In Section 4, two experiments have been conducted to evaluate the proposed framework by using a real-world dataset collected from Twitter. Section 5 concludes this study, as well as the limitations and future work.

Contact: Dr Weihua Li, Auckland University of Technology, 55 Wellesley St E, Auckland, 1010, NZ, +64 09921 9999

2. Related Work

Privacy leakage detection in OSNs has attracted great attention to many researchers. A few studies have been conducted to analyze user privacy revealing on Twitter. People are very cautious about their personal information, e.g., home address, phone number, etc., but they consciously or unconsciously disclose their plans and activities through posting information in OSNs [Humphreys 10]. Publishing such messages online can possibly raise serious security issues. For example, a message saying "going out for holiday" implies that no one stays at home, which may cause robbery. Therefore, users should be reminded before delivering such event-related information. Mao, Shuai and Kapadia (2011) present a detection approach to analyzing three types of sensitive tweets, i.e., drunk, vacation and disease tweets. The research on privacy issues is not restricted to Twitter. Acquisti and Gross (2006) investigate the privacy concerns of users on Facebook. Dwyer, Hiltz, and Passerini (2007) compare the trust and privacy issues between Myspace and Facebook. Bhagat, Cormode, Srivastava, and Krishnamurthy (2010) show that privacy can be revealed by predicted social graph.

Whereas, very few studies investigate how to detect individual privacy information and protect individual OSNs users from online privacy leak. Therefore, in this study, instead of assisting the organizations, we target the individual online users and keep them away from privacy leakage. As almost all the posts by users are unstructured data, the information extraction plays a pivotal role in the proposed framework.

Named Entity Recognition (NER) is an important method for extracting domain-specific information [Nadeau 07]. Given the context of privacy detection domain, NER can assists users in identifying privacy-related entities after given sufficient training. Traditionally, Conditional Random Field (CRF) classifier has been employed for NER due to its robustness and reliability. Gomez-Hidalgoy et al. (2010) proposed a mechanism which is capable of detecting named entities, e.g., a company, brand, or person, using NER. Nowadays, Bi-directional Long Short-Term Memory with Conditional Random Field (Bi-LSTM CRF) model becomes more popular as it achieves more promising results [Lample 16]. Therefore, we utilize Bi-LSTM CRF for privacy-related entities extraction. Privacy Information Detection.

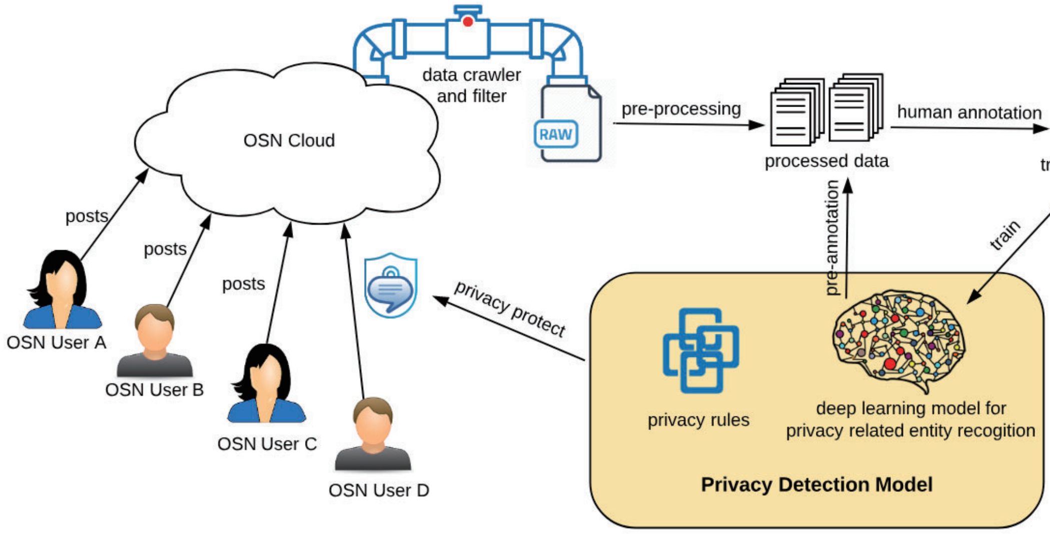

Fig 1 The Proposed Privacy Detection Framework

3. Privacy Information Detection Framework

3.1 Definition of Privacy

"Privacy" has a very broad meaning, which generally refers to the people's right to keep their personal matters and relationships secret [Gehrke 11]. In this sense, the privacy information is associated with something personal, as well as the matters of past, present, and future. Given the generic definition, in this paper, the privacy information that users tend to publish on the OSNs is defined as a sequence of words, stating or implying any individual's personal information, preferences, events that he or she involved.

Based on the definition given above, the privacy information incorporates four categories of entities, i.e., PERSON, TRAIT, PREF, and EVENT. More specifically, PERSON refers to any expression that identifies a real person; TRAIT represents the personally identifiable information, such as birth date and phone number; PREF refers to an individual's preference or hobbies; EVENT indicates the matters or activities that one involves anytime anywhere. Therefore, given a word sequence, the judgment of privacy information can be summarized as a rule as follows:

$$\exists \mathrm{PERSON} \wedge (\exists \mathrm{TRAIT} \vee \exists \mathrm{PREF} \vee \exists \mathrm{EVENT}) \longrightarrow \mathrm{P}(\mathrm{message})$$

3.2 The Automated Privacy Detection Framework

Our automated privacy information detection framework is demonstrated in Figure 1. The framework illustrates how the privacy detection model gets trained and utilized.

Users keep posting messages to the OSNs hosting in the cloud. Such raw unstructured and public data can be obtained through crawlers or APIs provided by the OSNs. For example, Twitter allows developers to search public tweets if the proposed project is approved. Given the context of privacy detection, the potential privacy-related data should be filtered and downloaded. Pre-processing is conducted based on specific rules, such as removing meaningless words and characters and parsing word sequences to tokens. The processed data are supposed to be further enriched by running through the pre-annotation if a privacy-detection NER model is available. Next, human involvements, i.e., manual annotation, are required. Specifically, according to the aforementioned definition of privacy information, it is essential to recognize the privacy-related entities, i.e., PERSON, TRAIT, PREF, and EVENT. The annotation also aims to figure out these four types of entities from the processed data. The annotated dataset is then fed into the deep learning model for training.

The privacy detection model consists of two components, i.e., a deep learning model for privacy-related entities recognition and privacy definition rules. For any posting messages by the OSN users, the proposed model is capable of judging whether the message is privacy-related or not. Moreover, as the privacy rules are properly defined, the privacy detection model can also explain the reason why the message is potentially privacy-related. Using a single deep learning model for private messages classification definitely loses the capability of justification.

3.3 Privacy-Related Entities Recognition

The privacy-related entities recognition plays an important role in the entire framework. There are two major aspects affecting the performance of an NER model, i.e., the annotation approach and the algorithm.

In this study, the Bi-LSTM CRF model has been employed for privacy-related entities recognition in our model, as it is capable of achieving more promising results compared with that of other classic algorithms when being applied to NER [Lample 16]. Bi-LSTM can learn long-term dependency due to the structure of the 'cell' in the hidden layer. Moreover, it can adjust the impact of previous states on the current states through the forget gate, input gate and output gate in the 'cell' [Graves 05]. However, it lacks the feature analysis on the sentence level, which can be solved by CRF. It can consider contextual conditions to make global optimal predictions. Combining the LSTM and CRF together can label sequence effectively when ensures to extract contextual features [Huang 15].

In regards to the annotation approach, BIO encoding scheme is utilized to tag entities in NER task [Kim 04]. BIO encoding scheme is a standard method which can solve the joint segmentation problem in labelling sequence by transforming them into raw labelling problem. Specifically, 'B-' is used as a prefix of an entity, implying the beginning of an entity; prefix 'I-' tags other characters indicating the tag is inside of an entity and 'O' is used for characters which do not belong to any pre-defined entities. For example, privacy-related entities fall into BIO scheme are normally annotated as follows:

I	*watch*	*a*	*movie*	*with*	*Christine.*
B-PERSON	**B-EVENT**	**I-EVENT**	**I-EVENT**	**O**	**B-PERSON**

4. Experiments

Two experiments have been conducted to evaluate the proposed privacy detection framework. The first experiment aims to train a privacy-related entities recognition model using Bi-LSTM CRF model. The second experiment gives some case studies to further demonstrate the effectiveness of the proposed privacy detection model.

4.1 Data description

Twitter is one of the largest OSNs, which enables users to conduct online social activities, including the distribution of any ideas or information. In Twitter, the messages that are posted and interacted by users are known as "tweets". Twitter provides APIs, allowing developers to search and store tweets. Therefore, we utilize Twitter API to collect 18k tweets by searching for some terms which potentially result in privacy leakage, such as pronouns, sensitive words, plans, etc.

4.2 Experiment 1

In Experiment 1, a privacy detection model based on Bi-LSTM CRF is trained to recognize the privacy-related entities. Through which, the users can be prompted before potential privacy leakage occurs. According to the definition of privacy and BIO encoding scheme mentioned previously, nine tags have been defined, i.e., 'B-PERSON', 'I-PERSON', 'B-TRAIT', 'I-TRAIT', 'B-PERF', 'I-PERF', 'B-EVENT', 'I-EVENT' and 'O'. Around 200 tweets have been annotated manually by applying these nine tags.

In this experiment, we leverage three traditional evaluation metrics as follows:

- **Precision**: the percentage that privacy-related entities can be labelled correctly among all the entities which are labelled privately in the test dataset.
- **Recall**: the percentage that privacy-related entities can be labelled correctly among all the actual privacy entities in the test dataset.
- **F1-score**: the weighted average of precision and recall, which takes both the two measures into account.

After 50 epochs' training, the performance of the deep learning model is demonstrated in Table 1.

Table 1 Performance of Privacy-Related Entities Recognition

Entity	Precision	Recall	F1-score
PREF	0.99	0.67	0.8
TRAIT	0.68	0.88	0.77
PERSON	0.98	0.93	0.95
EVENT	0.77	0.67	0.71
Avg/Total	**0.86**	**0.83**	**0.84**

4.3 Experiment 2 Case Study

In this experiment, we further demonstrate the effectiveness of the proposed privacy detection model by selecting three tweets posted recently and analyzing the results produced by the model.

Case 1: Adam and I are having lunch tomorrow.

Results: Adam (B-PERSON) and I (B-PERSON) are having (B-EVENT) lunch (I-EVENT) tomorrow.

Explanations: Based on the privacy rules, this tweet is privacy-related since it mentions both PERSON and EVENT.

Case 2: Watching a movie is a good way to relax!

Results: Watching (B-EVENT) a (I-EVENT) movie (I-EVENT) is a good way to relax!

Explanations: This tweet is just a simple statement regarding "Watching a movie", which is not a private one.

Case 3: My son is crazy about coke.

Results: My (B-PERSON) son (I-PERSON) is crazy about coke (B-PREF).

Explanations: This tweet talks about PERSON and PREF, it is privacy-related.

5. Conclusion and Future Work

In this paper, we presented a privacy information detection framework for individual OSN users. The objective is to protect end users from potential privacy leakage before posting any messages. The proposed framework explains the process of data collection, processing, model training and how it works. Both privacy rules and Bi-LSTM CRF model are leveraged in the privacy detection model. Thus, the proposed model is equipped with the capability of both detection and results explanation.

This study is still very preliminary and there is huge space for further investigation and extension. In the future, we intend to utilize a larger training dataset for performance evaluation and improve the performance of the privacy-related entities recognition by fine-tuning the parameters of Bi-LSTM CRF. Moreover, different tweets are associated with different degrees of privacy leakage. How to evaluate and score the privacy-leakage degree is also under our consideration.

References

[Wang 11] Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. "I regretted the minute I pressed share: A qualitative study of regrets on Facebook.". *Proceedings of the seventh symposium on usable privacy and security*,pp.10 (2011).

[Humphreys 10] Humphreys, L., Gill, P., & Krishnamurthy, B. "How much is too much? Privacy issues on Twitter." *Conference of International Communication Association* (2010).

[Mao 11] Mao, H., Shuai, X., & Kapadia, A. "Loose tweets: an analysis of privacy leaks on Twitter." *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society.* (2011).

[Hasan 13] Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. "A discussion of privacy challenges in user profiling with big data techniques: The excess use case." *Big Data (BigData Congress), 2013 IEEE International Congress on.* (2013).

[Wang 17] Wang, Q., Bhandal, J., Huang, S., & Luo, B. "Classification of private tweets using tweet content." *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on.*(2017).

[Aborisade 18] Aborisade, O., & Anwar, M. "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers." *2018 IEEE International Conference on Information Reuse and Integration (IRI).* (2018).

[Lample 16] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. "Neural architectures for named entity recognition." *arXiv preprint arXiv:1603.01360* (2016).

[Bengio 94] Bengio, Y., Simard, P., & Frasconi, P.. "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks* Vol.5, No.2, pp.157-166 (1994).

[Graves 05] Graves, A., & Schmidhuber, J. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." *Neural Networks* Vol.18, No.5-6, pp. 602-610 (2005).

[Huang 15] Huang, Z., Xu, W., & Yu, K. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991* (2015).

[Bhagat 10] Bhagat, S., Cormode, G., Srivastava, D., & Krishnamurthy, B. "Prediction Promotes Privacy in Dynamic Social Networks." *WOSN*. (2010)

[Acquisti 06] Acquisti, A., & Gross, R. "Imagined communities: Awareness, information sharing, and privacy on the Facebook." *International workshop on privacy enhancing technologies.* (2006)

[Dwyer 07] Dwyer, C., Hiltz, S., & Passerini, K. "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace." *AMCIS 2007 proceedings,* pp. 339 (2007).

[Kim 04] Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N."Introduction to the bio-entity recognition task at JNLPBA." *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications.* pp.70-75 (2004).

[Wu 15] Wu, Y., Xu, J., Jiang, M., Zhang, Y., & Xu, H. "A study of neural word embeddings for named entity recognition in clinical text." *AMIA Annual Symposium Proceedings*, Vol. 2015, p. 1326 (2015).

[Nadeau 07] Nadeau, D., & Sekine, S. "A survey of named entity recognition and classification." *Lingvisticae Investigationes*, Vol.30, No.1, pp.3-26. (2007).

[Gehrke 11] Gehrke, J., Lui, E., & Pass, R. "Towards privacy for social networks: A zero-knowledge based definition of privacy." *Theory of Cryptography Conference*, pp. 432-449 (2011).

[Gomez-Hidalgo 10] Gomez-Hidalgo, J. M., Martin-Abreu, J. M., Nieves, J., Santos, I., Brezo, F., & Bringas, P. G. (2010, August). Data leak prevention through named entity recognition. In Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 1129-1134). IEEE.