A Tool for Automatic Assessment and Awareness of Privacy Disclosure

Paolo Cappellari
City University of New York
College of Staten Island
2800 Victory Blvd
Staten Island, New York 10314
paolo.cappellari@csi.cuny.edu

Soon Ae Chun
City University of New York
College of Staten Island
2800 Victory Blvd
Staten Island, New York 10314
soon.chun@csi.cuny.edu

Mark Perelman
City University of New York
College of Staten Island
2800 Victory Blvd
Staten Island, New York 10314
mark.perelman@cix.csi.cuny.edu

ABSTRACT

With increasing frequency, the communication between citizens and institutions occurs via some type of e-mechanism, such as websites, emails, and social media. In particular, social media platforms are widely being adopted because of their simplicity of use, the large user base, and their high pervasiveness. One concern is that users may disclose sensitive information beyond the scope of the interaction with the institutions, not realizing that such data remains on these platforms. While awareness about basic data (e.g. address, date of birth) protection has risen in the past few years, many users still neglect or fail to realize the amount and significance of the personal information deliberately or involuntarily disclosed on these communication platforms. Determining private from non-private data is difficult. The goal of this work is to devise a method to detect messages carrying sensitive information from those that not. Specifically, we employ machine learning methods to build a privacy decision making tool. This work will contribute to develop a privacy protection framework where a client-side privacy awareness mechanism can alert users of the potential private information leakages in their communications.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Applied computing → E-government**;

KEYWORDS

Social media, e-government, supervised learning, Twitter

ACM Reference format:
Paolo Cappellari, Soon Ae Chun, and Mark Perelman. 2017. A Tool for Automatic Assessment and Awareness of Privacy Disclosure. In *Proceedings of dg.o '17, Staten Island, NY, USA, June 07-09, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3085228.3085259

1 INTRODUCTION

Governments across many countries are embracing social media as a channel to interact with citizens, e.g. [2, 3, 8]. Over the past few years, users have become accustomed to share a vast variety of information on social networks, including personal information. This habit may leak in their interactions with the institutions, therefore providing more, potentially sensitive, information than necessary. As a matter of facts, more and more users are directly or indirectly sharing information about themselves or others, e.g. friends [4, 5, 7].

Users are more aware than in the past that organizations monitor their online behaviors for multiple purposes, including market segmentation, customer profiling, target advertisement, etc. Users have become more wise in sharing (or not) personally identifiable information (PII), which are rarely found on public spaces. On the other hand, it has been observed that there is a whole different corpora of data that users share without much attention [9] when interacting on social networks. It seems that users do not think of social posts as a mean to share sensitive information. Privacy is one of grand challenges in the Web 2.0 social web era with mass participation and sharing data through online socialization is a norm, and yet it is one of the most elusive topics to be studied scientifically. This project wants to study the issue of sensitive information disclosure, also referred to as information leakage, from a user awareness perspective. The ultimate goal is to build a platform to raise awareness and educate users about information leakage. The problem is further complicated by the fact that information leakage is, to some extent, a subjective matter: some information is regarded as private by some individuals, but not from others.

Information leakage is becoming a topical area of research, as more researchers are trying to tackle the problem from different perspectives. For example, in [7] authors propose an automatic system to classify Twitter posts in a few, fixed, categories. The work in [1] analyze multiple privacy issues related to the use of Facebook social network, but does not help users in preventing information leakage on their messages. Other works such as [6], try to raise privacy awareness by prompting users whenever some sensitive information may be disclosed during web navigation, but limits its scope of action to basic, structured, data such as email, name, social status. Overall, a comprehensive approach helping users assessing and overcoming information disclosure issues is missing.

With this work we present an approach to automatically detect messages carrying private information. Our main goal is to detect message carrying private information so that to alert the user before the message itself is posted on a social platform. In the next sections we describe our approach, we presents our experiments, and draw a line for future work.

dg.o '17, Staten Island, NY, USA

DOI: http://dx.doi.org/10.1145/3085228.3085259

SVM	Non-private	Private	Precision
Predicted/Not-Private	55	41	57.29%
Predicted Private	45	100	68.97%
Recall	55.00%	70.92%	

Table 1: Automatic privacy disclosure assessment via SVM.

Rule Induction	Non-private	Private	Precision
Predicted/Not-Private	55	55	50.00%
Predicted Private	45	86	65.65%
Recall	55.00%	60.99%	

Table 2: Automatic privacy disclosure assessment via Rule Induction.

2 APPROACH

We consider a supervised machine learning problem and train classifiers on individual tweets of a generic social media user to predict whether the content of the post lean is disclosing private information or not. A sample of 500 tweets has been retrieved from the Twitter sample stream to construct a dataset for the machine learning model. These tweets have been manually annotated with labels private and not-private. The annotation process has been crowd-sourced by surveying the opinion of about 100 volunteers. Volunteers were presented with about 30 tweets each. In turns, each tweet has been presented to roughly 6 different volunteers. We developed a basic web application where the volunteers had to read one tweet at the time and decide whether to deem it as private or not-private Eventually, all survey data is collected and each tweet is associated with a final privacy label, that is the annotation that has received the majority of votes. Before training the predictive model to assess tweets' privacy, the labeled tweets are preprocessed to: remove common and stop words, replace each word with a common synonym, via the Wordnet lexical database; and each word is stemmed to remove derived or inflected variations so to reduce the dictionary of terms to words in their root (or base) form.

The refined set of labeled tweets is used to train and verify the outcomes of several machine learning classifiers. The refined dataset has been split in two sets: 80% was used to train the model, and the remaining 20% to validate the results. We have experimented the following classifiers: Nearest Neighbor, Rule Induction, Random Forest, Naive Bayes, and multiple variations of Support Vector Machine (SVM). These models are available in RapidMiner, which is the tool we used to conduct our experiments.

3 RESULTS AND CONCLUSION

We tested multiple machine learning algorithms, with the goal of being able to assess whether the text composing the social post is disclosing sensible information or not. The best performing supervised learning method resulted to be SVM, for which the results are shown in Table 1, followed by the Rule Induction, see Table 2, and Naive Bayes, see Table 3.

The results are encouraging although more work is required to improve the quality of the prediction assessments. In the best case,

Naive Bayes	Non-private	Private	Precision
Predicted/Not-Private	55	62	47.00%
Predicted Private	45	79	63.71%
Recall	55.00%	56.03%	

Table 3: Automatic privacy disclosure assessment via Naive Bayes.

precision ranges from 60% to 70% and recall from 55% to 70%. All models are consistent in finding the not-private tweets, while they exhibit variations in identifying the private ones. This performance can likely be improved by considering a larger training dataset. One issue, however, is building a large, unbiased, training dataset. Most approaches in other works rely on a synthetic way of constructing the training set, where depending on what keyword(s) occurs in the text, a message is automatically annotated as private (or not). These works neglect to consider that the same term (keyword) may assume different meaning depending on the context. We plan to address this and other issues in future work.

REFERENCES

[1] Alessandro Acquisti and Ralph Gross. 2006. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *Proceedings of the 6th International Conference on Privacy Enhancing Technologies (PET'06)*. Springer-Verlag, Berlin, Heidelberg, 36–58. DOI: https://doi.org/10.1007/11957454_3

[2] Lori A. Brainard and John G. McNutt. 2010. Virtual GovernmentâĂŞCitizen Relations. *Administration & Society* 42, 7 (2010), 836–858. DOI: https://doi.org/10.1177/0095399710386308

[3] Jeremy Crump. 2011. What Are the Police Doing on Twitter? Social Media, the Police and the Public. *Policy & Internet* 3, 4 (2011), 1–27. DOI: https://doi.org/10.2202/1944-2866.1130

[4] Balachander Krishnamurthy and Craig Wills. 2009. Privacy Diffusion on the Web: A Longitudinal Perspective. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 541–550. DOI: https://doi.org/10.1145/1526709.1526782

[5] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. 2013. Privacy Awareness About Information Leakage: Who Knows What About Me?. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES '13)*. ACM, New York, NY, USA, 279–284. DOI: https://doi.org/10.1145/2517840.2517868

[6] Delfina Malandrino and Vittorio Scarano. 2013. Privacy leakage on the Web: Diffusion and countermeasures. *Computer Networks* 57, 14 (2013), 2833 – 2855. DOI: https://doi.org/10.1016/j.comnet.2013.06.013

[7] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (WPES '11)*. ACM, New York, NY, USA, 1–12. DOI: https://doi.org/10.1145/2046556.2046558

[8] Jesper SchlÃęger and Min Jiang. 2014. Official microblogging and social management by local governments in China. *China Information* 28, 2 (2014), 189–213. DOI: https://doi.org/10.1177/0920203X14533901 arXiv:http://dx.doi.org/10.1177/0920203X14533901

[9] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*. ACM, New York, NY, USA, Article 10, 16 pages. DOI: https://doi.org/10.1145/2078827.2078841