Survey of Sensitive Information Detection Techniques: The need and usefulness of Machine Learning Techniques

Riya Shah
Indus University, Ahmedabad

Manisha Valera
Indus University, Ahmedabad

ABSTRACT

The amount of digital data generated is growing by the day and so is the need to protect sensitive content from being published on the Internet. This paper talks about studies in the field of content-based detection of both personal sensitive information like name, birth date, medical records and corporate sensitive information like confidential agreements, earnings reports. It describes previous dominant technology in data leakage prevention along with their shortcomings, resulting in the need for machine learning being used. Since personal sensitive information usually has a certain pattern, regular expressions are the most efficient way to detect those. However, corporate sensitive information generally does not follow a particular pattern and varies in different companies. This paper talks about research related to the use of machine learning in detecting sensitive content, eventually concluding the need for further research in detecting corporate sensitive information.

KEYWORDS

Data leakage prevention, Machine learning, Supervised learning, Sensitive information

INTRODUCTION

Leakage of sensitive information is an issue whose seriousness has escalated over the years given the increase in usage of social media as well as the storage of huge amounts of sensitive data by companies. Sensitive information not only involves a company's confidential documents but also includes any personally identifiable information such as name, birth date, medical records [1]. While some of this is intentionally placed on the Internet by the person it rightfully belongs to, a large amount of sensitive information also ends up being intentionally or unintentionally leaked. The problems caused by these leakages are not even enumerable; they could range from identity thefts to million dollar losses.

Previously, the dominant methods of content-based detection of private information were pattern matching, fingerprinting and natural language analysis [2]. Pattern matching involves finding particular keywords with the use of regular expressions. However, this limits the type of sensitive information that can be detected. For example, pattern matching may detect a credit card number or Social Security number but is unable to classify the press release statement of a company as "not sensitive" and a customer master of the same company as "sensitive", considering they both contain the keyword they are looking for. For fingerprinting, each sensitive file must be fingerprinted (a hashing algorithm is used to convert the sensitive input text to a hash value) before it can be detected as sensitive by the algorithm. This requires that each sensitive file be found first and then be fingerprinted. This could be a very lengthy and tiring process. Thus, a lot of studies focus on finding more efficient methods to reduce leakage - with a growing interest in the usage of machine learning; specifically, supervised learning methods.

LITERATURE REVIEW

Twitter, being a widely used social media platform, has been the focus of most research for detecting leaks of personal sensitive information. In [3], they analysed tweets to detect sensitive information revealed by the users. First, the collected tweets were categorized into: vacation tweets, drunk tweets and disease tweets. For each category of tweets, a dataset of "sensitive" and "not sensitive" tweets was made and labelled. For instance, in the case of vacation tweets, "sensitive" tweets are those that reveal concrete vacation plans while the rest are "not sensitive". This dataset was then used to train a classifier using machine learning algorithms such as Naive Bayes and SVM. However, better results were obtained using Naive Bayes. This is surprising since most text classification tasks favour SVM [4]. The resulting classifier was then used to detect unlabelled sensitive or not sensitive tweets of a particular category it was trained for. They also describe how selecting different features, improves or degrades their results, thus emphasizing its importance in any machine learning problem. Despite the fact that machine learning algorithms can be applied to sensitive information detection, it is also subtly stated that this is only feasible when the problem domain is carefully narrowed down by categorizing tweets (in this case, into vacation, drunk and diseases tweets) and building datasets of sensitive and not sensitive tweets for each category. If the categorization step is removed, it would make it extremely difficult to classify all the tweets.

In [5], they also used data from Twitter to detect private information leaks. However, their detection is more generalized than Mao's. They used topic modelling (Latent Dirchlet Allocation to find different topics in the given data), privacy ontology (a 'privacy dictionary' tool for performing automated private information detection [6]), named entity recognition (to detect names, locations, company or organization names) and sentiment analysis (to label each sentence as private or not private). For each users' tweets, they classified them as containing or not containing private information, and assigned a privacy score (1,2 or 3) to each user depending on the percentage of their tweets that contain private information. They picked 9 categories of private information and picked their features accordingly. The machine learning algorithms used were SVM, Naive Bayes and AdaBoost, with AdaBoost performing best. These algorithms were trained on the features obtained and the resulting classifier predicted the privacy score of a new user. However, it only reached an accuracy of 69.63%. One of the reasons for this could be that the training dataset was annotated by AMT workers and not by the people who actually wrote the tweets. This could lead to tweets being given labels that are not the same as what the writer of the tweet would label them. Thus, given a carefully labelled dataset, classification of private and not private information may be easier using machine learning algorithms. Like [3], [5] also emphasizes the need to categorize private information before machine learning can be used.

Unlike the aforementioned studies, [7] focused on the usage of corpus based association rules instead of machine learning techniques. Assuming that a company knows a keyword related to which it must detect sensitive information, their method proves extremely beneficial as it narrows the search down. They use association rule mining (finding frequent patterns) to detect words that occur frequently together in a dataset. Therefore, given a keyword by the company, the algorithm searches the company's given training dataset to find words that most frequently co-occur with the keyword. The paper gives a great example - if Apple was to look for leakage of sensitive content regarding the release of its new SIM feature, it could find all the words that co-occur with SIM in its training corpus and then search the input text under consideration for those words to limit their search. The input texts under scrutiny are then searched for all those words. This method is particularly useful when the presence of a word or words that co-occur in sensitive information is definite. Therefore, this method

can be applied as a pre-processing step to decrease the amount of information that must be searched for sensitive content.

Despite the number of studies performed for finding personal sensitive content, I did not find a lot of resources for finding corporate sensitive content. The reason for this could be the unavailability of a proper dataset to work on as outlined by Nguyen in [8]. Companies would not be willing to share their sensitive information for the purpose of this research and a public data set that already contains sensitive information is not available. Nguyen's research suffered due to the above mentioned reasons, causing him to generalize his research to only the classification of corporate and public text. The corporate text could either be sensitive or not. However, he does mention that SVM turned out to be the best contender for text classification and this suggests that it would perform well in the task of classifying corporate private and corporate public text as well.

Symantec was the first company to introduce machine learning as part of its data leakage protection software [9]. They called it vector machine learning and they've provided a great deal of detail about the kinds of data it can be trained with and how it works [10]. The program requires positive and negative samples of sensitive data from a particular company. A profile is built using the samples as training data. The accuracy of the profile is presented to the user who may then select if the profile is run on their network to detect similar sensitive data or not. If the profile is run, any unknown document is assigned a similarity score between 0.0 and 10.0. A score of zero would mean that it is not at all similar to any of the training documents while 10.0 would mean it is exactly like one of the training documents [2]. However, there is no mention of which machine learning technique specifically is being applied to their software.

CONCLUSION

The limitations of keyword search, regular expression matching and fingerprinting can be overcome using machine learning techniques. For example, all the methods mentioned above can be paired with supervised learning methods to build stronger data leakage prevention software. In the case of detecting personally identifiable information like birth date or credit card number or SSN, regular expression matching would work well while for classifying corporate documents to private/confidential and public, supervised learning methods would work well. Thus, the usage of machine learning depends on the characteristics of the sensitive content that is being found.

Given the lack of research in the field of the efficiency of various supervised learning methods like SVM and Naive Bayes, it would be a good idea to collect a dataset of actual labelled corporate documents and study the performance of various supervised learning methods in classifying those documents. Usage of machine learning would also be advantageous as the resulting classifiers would depend on the training set its given. Thus, companies would have the liberty to tweak the performance of the classifier by changing the training set. Further research in this field would be extremely helpful.

BIBLIOGRAPHY

[1] "U.S. General Services Administration," [Online]. Available: http://www.gsa.gov/portal/content/104256. [Accessed February 2016].

[2] "Symantec," [Online]. Available: eval.symantec.com/mktginfo/enterprise/white_papers/b-dlp_machine_learning.WP_en-us.pdf. [Accessed January 2016].

[3] H. Mao, X. Shuai and A. Kapadia, "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)*, Chicago, 2011.

[4] T. Joachims, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features," in *Proceedings of the 10th European Conference on Machine Learning (ECML '98)*, 1998.

[5] A. Caliskan-Islam, J. Walsh and R. Greenstadt, "Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES'14)*, Scottsdale, 2014.

[6] A. J. Gill, A. Vasalou, C. Papoutsi and A. Joinson, "Privacy dictionary: a linguistic taxonomy of privacy for content analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11) Pages 3227-3236*, Vancouver, 2011.

[7] R. Chow, P. Golle and J. Staddon, "Detecting Privacy Leaks using Corpus-based Association Rules," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08) Pages 893-901*, Las Vegas, 2008.

[8] K. Nguyen, *Classification of Corporate and Public Text*, 2011.

[9] "Information Week," [Online]. Available: http://www.darkreading.com/risk/symantec-integrates-machine-learning-into-dlp/d/d-id/1134931. [Accessed 2016].

[10] "Symantec," [Online]. Available: https://support.symantec.com/en_US/article.TECH219962.html. [Accessed 2016].