Proposal for Master Thesis

DV2566: Master Thesis in Computer Science

Version Number – February 7, 2016

Thesis	Tentative title	Text classification of Legitimate and Rogue online Privacy Policies
	Classification	Privacy policies, clustering algorithms, statistical testing, text classification
Student 1	Name	Kaavya Rekanar
	e-Mail	kare@student.bth.se
	Social security nr	940521-7184
Supervisor	Name	Martin Boldt
	e-Mail	martin.boldt@bth.se
	Department	Department of Computer Science
External	Name and title	
	e-Mail	

1. Introduction

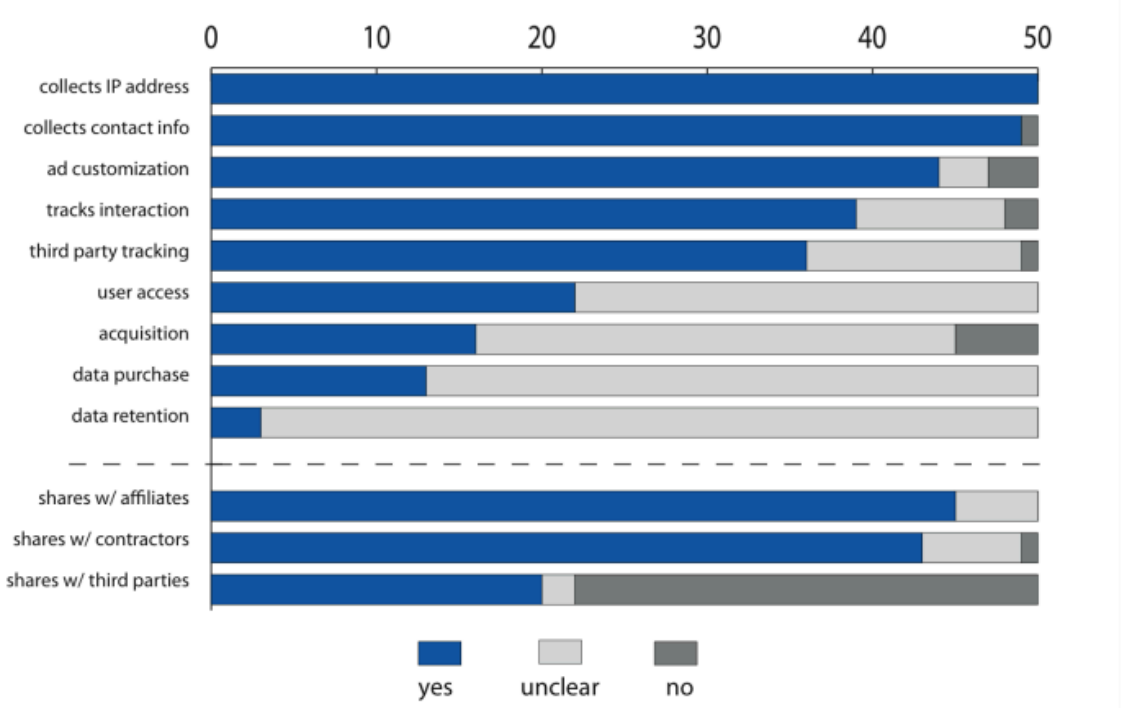
A privacy policy is an affidavit that reveals the working of a certain firm that manoeuvers and operate a website that handle various amounts of client data, e.g., cookie files, etc. On one hand, these policies enable users to engage in transactions and interactions on the Internet, while on the other hand, abuses and leakage of this information could violate the privacy of their owners, sometimes leading to serious circumstances [1].
A privacy policy is always a matter of trust. Studies reveal that an estimation of 77% companies are putting up a privacy policy in recent times [2]. These policies differ greatly from one institution to another and they address many issues that are different than those the users care about [2].
The Federal Trade Commission has set up some rules in designing of a privacy polices, which if violated lead to the formation of a rogue privacy policy, that could possibly cause harm to a user [3]. A legitimate policy is the one, which follows all the rules set up the Commission.
A rogue policy has the ability to interfere with a user's privacy, which could lead to pretty sticky circumstances one would want to avoid if possible. Hence, the differences between legitimate and rogue privacy policies can be listed out and an ability to easily differentiate them using some machine learning algorithms would be a good start to save all the trouble to a user. This project will investigate the possibility to separate between legitimate and rogue privacy policies using supervised machine-learning algorithms.

Related Work

In 2003, the Online Privacy Protection Act has been enacted which requires website owners to post a statement of their policies regarding the collection and sharing of personal information in the California State Legislature [3]. Though the goal of this legislation was to create some transparency about the data collection practices and to help users make informed decisions, it does not regulate the substance of websites' practices [3]; they only need to disclose those practices [3].

Privacy policies have been rendered as ineffective due to several reasons:
First, due to the fact that Privacy Policies are difficult to read- Most of them are written in legal jargon that makes it difficult for an average person to read and understand [2][13], because of which most of them do not bother to read them [2].
The framing of privacy policies is such that, most of them lead customers to believe that their privacy is protected and concerned for [12]. A study found that "Users do not read privacy policies because they believe that they do not have to; to consumers, the mere presence of a privacy policy implies some level of often false privacy protection" [12].
Even if a consumer can understand the privacy policy, they are not interested to invest the amount of time required to read privacy policies [2][7]. A study has proved that it would take an average person about 200 hours a year [7] to actually read the policy for every unique website visited in a year, not to mention the updated version of policies for sites visited on a repeated basis [7].
Even if they could understand the policy and make time to do so, there is not enough market differentiation for users to make informed choices [14]. Furthermore, many website policies are vague about what user information they are collecting and how it is going to be used. As they are all equally poor, the users have no viable alternatives [14]. This is a market failure.
Lastly, even if there was a market differentiation, it is not comprehensible that the users will protect themselves [11]. The potential danger are not salient [9], not to mention the fact that they are difficult to evaluate against the benefits [10] of using a website [9][10][11].

Rating the privacy policies in three categories: Yes, Unclear and No based on the information the privacy policies collect from the users, this graph shows the data collection mechanism of 50 websites whose policies have been already searched [11]. The privacy policies have been collected from the top-50 corporations on the Forbes 500 list.



The Platform for Privacy Preferences Project (P3P) enables websites to express their privacy practices in an XML-based machine readable format than can be retrieved automatically and interpreted easily by user agents, by the means of P3P-enabled browsers [22]. As a first step, the website sends a machine-readable proposal of its privacy policies [22]. The proposal can be automatically parsed by a user agent and compared with the user's privacy preferences, thus the

users do not need to read the privacy policies of every website they visit [23]. Privacy Bird [24] and Privacy Finder [25] are examples of P3P user agents, able to compare P3P policies with user preferences. A limitation of the P3P policies is with users preferences. A limitation of the P3P is that it needs server-side adoption, which is not easily obtained: according to [26] only 20% of the websites amongst the E-Commerce Top 300 is P3P enabled [26]. Rogue actors defiantly won't assist users by adding P3P policies to their sites.

There has been no research particularly on rogue privacy policies as to my knowledge. But the paper, "A Machine Learning Solution to Assess Privacy Policy Completeness" [22] deals with a solution to show the legitimacy of a policy, which considered in the opposite direction, would make it a rogue policy; if according to the algorithm, the chosen one doesnot satisfy all the criteria to fit into a legitimate policy. The metrics that have been chosen in this literature are precision and recall to test the completeness of a policy. This is the knowledge gap that I identify and target to fill with this work.

According to [22], the k-NN classifier algorithm has been known to be one of the top-performing approaches in the text categorization tasks [22]. It has a minimal training stage and an intensive, time-consuming testing stage [22]. In the k-NN classifier, k is the number of closest neighbors, i.e., the training items considered [22]. But decision tree based approaches are more appealing for text learning [15], because their performance compares favorably with other learning techniques as well [15]. There are several classic decision tree algorithms, such as ID3 [17], C4.5 [18], and CART [16].

Non-linear SVM algorithms are useful when the gap between different categories cannot be linearly modeled and thus, a more complex function is needed [19][20]. There are many alternatives to the linear kernel used in the linear SVM; but the RBF kernel outperforms other variants due to the text classification tasks [20][21].

2. Aim and Objectives

The aim of this project is to find a working solution to discriminate rogue privacy policies from legitimate policies. This solution could be added to implement a client side browser add-on and later be developed as a complete software module for future work.

The objectives of this project are:

- Decide on which algorithms to include in comparison of the dataset.
- Decide on which metrics to use.
- Investigating which machine learning platform/suite to use.
- Collect dataset.
- Carry out experiment.
- Evaluate the feasibility of the experiment by testing it on the dataset previously collected.
- Investigate how to implement client side browser add-on.

3. Research Questions

The research questions to be answered in the thesis are:

1) To what extent can text classification algorithms distinguish the content in rogue privacy policies from legitimate?
2) What differences between legitimate and rogue privacy policies can be learnt from the evaluated classification algorithms?
3) How can a browser add-on be constructed based on the text classification results in RQ1?

4. Method

The proposed method of research is experiment, which makes use of one dataset, i.e., text containing 100 legitimate privacy policies and 50 rogue privacy policies. Rogue policies are a little difficult to collect, hence the reason for taking a smaller count. Using the guidelines provided by FTC, as to which qualities make a privacy policy a complete legitimate document, we can set up the metrics.

In this study, we will consider different evaluation metrics, e.g., precision, recall, AUC, F-measure and accuracy. The exact metrics used is decided as the review of different evaluation metrics is over and implemented accordingly.

The testing of the proposed system would be done using stratified 10-fold cross validation method, where the percentage of legitimate and rogue policies would be 66% and 33% respectively. This test can be done using machine learning suites like Weka, which makes the process faster.

Kruskal-Wallis test will be used for statistical testing. It is a non-parametric method for testing whether samples originate from the same distribution. This test can be used to compare two or more independent samples of equal or different sample sizes. Here, the metrics are dependent variables, and the algorithms chosen to compare and test would be independent variables.

After a significant difference is clearly identified between both the groups in the dataset, post-hoc test (one-way ANOVA) is run to confirm where the differences occurred.

Validity Threats

If the policies collected as rogue and legitimate really are not as they are claimed. Then, the whole study would be done thinking that they represent something they are actually not.

However, that threat can be overcome if the policies are manually read and compared with the FTC guidelines.

5. Expected Outcome

The proposed work should lead to validation and verification of privacy policies, so as to clearly differentiate the rogue from the legitimate ones.

This might lead to the development of software module, which would warn users from being manipulated by rogue policies in future.

6. Time and activity plan

A work breakdown structure has been created to show the working for thesis writing and project work. The time taken to complete the tasks and the dates has also been mentioned where important submissions are involved.

Task	Time	Date
Background Study	2 w	
Submit Proposal		07-02-2016
Collecting dataset and verifying against FTC rules	3w	
Decide which metrics to use	1w	
Investigate the platform to work on	1w	
Choose algorithms	2w	
Carry out the experiment	3w	
Evaluate the feasibility	3w	
Investigating implementation of client-side browser add-on.	2w	
Finalizing the thesis.	2w	18-05-2016

7. Risk Management

The management of risk to a maximum extent from the threats posed could be in the following manner.

- **Risk-** What if the chosen algorithms do not satisfy our requirements, and do not provide results as expected?
  The outcome may not be as expected, if the chosen algorithm to compare the dataset is a bad choice. That will also affect the metrics decided to use, hence the algorithm must be checked manually.
  Management- While choosing the algorithm and metrics, I have to be extra careful and recheck at every next step; to be sure that I do not waste spending time on a completely irrelevant concept not relating to my project.

- **Risk-** If there is any hardware crash, the whole experimental setup would be at a loss.
  Management- Continuous backups through out the project would be helpful in case of any hardware crash.

8. References

[1]. M. C. Mont, S. Pearson, and P. Bramhall, "Towards accountable management of identity and privacy: sticky policies and enforceable tracing services," in 14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings, 2003, pp. 377–382.

[2]. C. Jensen and C. Potts, "Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2004, pp. 471–478.

[3]. A.D.Miyazaki and S. Krishnamurthy, "Internet Seals of Approval: Effects on Online Privacy Policies and Consumer Perceptions," The Journal of Consumer Affairs, vol. 36, no. 1, pp. 28–49, 2002.

[4]. Kruskal; Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association **47** (260): 583–621.doi:10.1080/01621459.1952.10483441

[5]. Corder, Gregory W.; Foreman, Dale I. (2009). Nonparametric Statistics for Non-Statisticians. Hoboken: John Wiley & Sons. pp. 99–105.ISBN 9780470454619.

[6]. Siegel; Castellan (1988). Nonparametric Statistics for the Behavioral Sciences (Second ed.). New York: McGraw–Hill. ISBN 0070573573.

[7]. McDonald, Aleecia; Cranor, Lorrie Faith, ―The Cost of Reading Privacy Policies,‖ CyLab, Carnegie Mellon University, 2008.

[8]. Acquisti, Alessandro, ―Privacy in Electronic Commerce and the Economics of Immediate Gratification,‖ 2004. http://www.heinz.cmu.edu/~acquisti/papers/privacy-gratification.pdf

[9]. Acquisti, Alessandro; Grossklags, Jens, ―Privacy and Rationality,‖ Privacy and Technologies of Identity, 2006. http://www.dtc.umn.edu/weis2004/acquisti.pdf

[10]. Acquisti, Alessandro; Grossklags, Jens, ―What Can Behavioral Economics Teach Us About Privacy,‖ Digital Privacy: Theory, Technologies and Practices, 2007. http://www.heinz.cmu.edu/~acquisti/papers/Acquisti-Grossklags- Chapter-Etrics.pdf

[11]. Nehf, James, ―Shopping for Privacy Online,‖ Journal of Consumer Affairs, vol. 41, 2007. http://ssrn.com/abstract=1002398

[12]. Hoofnagle, Chris; King, Jennifer, ―What Californians Understand About Privacy Online,‖ Samuelson Law, Technology & Public Policy Clinic, 2008. http://www.law.berkeley.edu/clinics/samuelsonclinic/files/online_report_final.pdf

[13]. Anton, Annie, ―The Lack of Clarity in Financial Privacy Policies and the Need for

Standardization,‖ IEEE Security & Privacy, vol. 2, no. 2, 2004.
[14]. Cranor, Lorrie Faith, et al., "2006 Privacy Policy Trends Report,‖ CyLab Privacy Interest Group, 2007.
[15].C. Apt´e, F. Damerau, and S. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems (TOIS), 12(3):233–251, 1994.
[16]. L. Breiman. Classification and regression trees. Chapman & Hall/CRC, 1984.
[17]. J. Quinlan. Induction of decision trees. Machine learning, pages 81–106, 1986.
[18]. J. Quinlan. C4. 5: programs for machine learning. Morgan kaufmann, 1993.
[19]. G. Smits and E. Jordaan. Improved SVM regression using mixtures of kernels. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, volume 3, pages 2785–2790. IEEE, 2002.
[20]. J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. Neural processing letters, 9(3):293–300, 1999.
[21]. Z. Wang, Y. He, and M. Jiang. A comparison among three neural networks for text classification. In Signal Processing, 2006 8th International Conference on, volume 3, pages 1–4. IEEE, 2006.
[22]. E. Costante, Y. Sun, M. Petković, and J. den Hartog, "A Machine Learning Solution to Assess Privacy Policy Completeness: (Short Paper)," in *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, New York, NY, USA, 2012, pp. 91–96.
[23]. J. Reagle and L. Cranor, "The platform for privacy preferences: Communications of the ACM", 42(2):48–55, 1999.
[24]. L. Cranor and M. Arjula, "Use of a P3P user agent by early adopters". In the 2002 ACM workshop on Privacy in the Electronic Society, pages 1–10, 2002.
[25]. J. Tsai, S. Egelman, L. Cranor, and A. Acquisti, "The effect of online privacy information on purchasing behavior: An experimental study", Information Systems Research, 21(June), 2010.
[26]. P. Beatty, I. Reay, S. Dick, and J. Miller, "P3P Adoption on E-Commerce Web sites: A Survey and Analysis" IEEE Internet Computing, 11(2):65–71, Mar. 2007.