Automatic Categorization of Privacy Policies: A Pilot Study

Waleed Ammar[*] **Shomir Wilson**[†]
Norman Sadeh[†] **Noah A. Smith**[*]

December 2012
CMU-ISR-12-114
CMU-LTI-12-019

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]Language Technologies Institute, Carnegie Mellon University
[†]Institute for Software Research, Carnegie Mellon University

Abstract

Privacy policies are a nearly ubiquitous feature of websites and online services, and the contents of such policies are legally binding for users. However, the obtuse language and sheer length of most privacy policies tend to discourage users from reading them. We describe a pilot experiment to use automatic text categorization to answer simple categorical questions about privacy policies, as a first step toward developing automated or semi-automated methods to retrieve salient features from these policies. Our results tentatively demonstrate the feasibility of this approach for answering selected questions about privacy policies, suggesting that further work toward user-oriented analysis of these policies could be fruitful.

Keywords: Privacy, Natural Language Processing, Machine Learning

1 Introduction

Privacy policies for websites tend to be lengthy and difficult for the average internet user to read. Details of interest are often hidden in plain sight by verbose language [6], and even users who seek those details may come away confused and no better informed. At the same time, such policies are legally binding, and users of websites and online services are subject to them whether or not they are aware of their contents. Prior efforts, resulting in the W3C's P3P standard, have attempted to encourage website operators to provide their privacy policies in a precise, machine-readable format that can be processed by client-side software and displayed to the user in some relevant form. However, P3P has seen limited adoption [2], and even many sites that provide P3P policies have misused the specification [5]. It seems plausible that businesses, governments, and other institutions are not sufficiently motivated to make their privacy policies easier to understand.

This report describes a pilot experiment to estimate the extractability of salient features from website privacy policies. We experiment with automatic text categorization to answer some simple categorical questions about privacy policies. This is a preliminary stage in a larger research effort to use methods from natural language processing and machine learning to analyze the texts of privacy policies and retrieve salient features (e.g. those found in privacy nutrition labels described by Kelley et al. [4]) from them. This approach has the potential of requiring little (if any) cooperation on the part of website owners beyond posting a policy. It also has the potential to reduce the amount of effort that presently must be committed to crowdsource the extraction of salient features from privacy policies. The results of this pilot tentatively suggest the feasibility of this approach for selected common properties of privacy policies. Further efforts will determine whether this approach is practical for the full range of privacy nuances that internet users tend to care about, and how to present this information to users as they interact with websites.

2 Approach

2.1 Data

We used crowdsourced annotations for *privacy policy* and *terms of service* documents[1] of 57 websites from the "Terms of Service; Didn't Read" project (`http://tos-dr.info`). For each website, annotators identified a number of noteworthy terms of these documents governing use of the site's services, and gave brief textual descriptions for them such as:

- Deleted images are not really deleted
- Using your real name is optional
- Notifications [of a change in policy] 30 days before changes [take effect]

[1] For simplicity, the term *privacy policy* will refer to the union of both of these kinds of documents throughout the paper.

The set of descriptions is essentially open; in fact half of the descriptions are only used once (i.e., to annotate a single document). A few concepts are repeated (often with rephrased descriptions) multiple times across documents, and some capture concepts related to privacy. The most common concepts in the data are:

- Ability to leave the service (found in 21 policies)
- Transparency on law enforcement requests (found in 19 policies)
- Providing a notice before changing the terms (found in 10 policies)

In addition to the annotated privacy policies, we also collected 794 privacy policies for which we did not have any annotations. Before feature extraction, we preprocessed the privacy policy documents by lowercasing the text and removing punctuation and stopwords.[2]

2.2 Model

We use logistic regression, a classic high-performance probabilistic model, to map privacy policy documents to categorical labels. In this pilot study, there are two labels, corresponding to presence and absence of a concept. The function is a linear classifier:

1. A document d is converted into a vector representation, with each dimension corresponding to unigrams, bigrams and trigrams (i.e., sequences of up to three consecutive words), and the value denoting term frequency in d. We call this vector x_d.
2. A scalar score is calculated as the inner product of x_d and an optimized weight vector w. We discuss how the weights are optimized below.
3. If the score is above a threshold, we hypothesize that the concept is present. Otherwise, the concept is hypothesized to be absent.

We start by training a model on labeled instances. Training is accomplished by solving a convex optimization problem with respect to the model weights w. We use stochastic gradient descent to minimize the average L_2-regularized log-loss of the model. We use convergence of the log-likelihood of the training set to determine when to stop the iterative training process. The log-likelihood converges when the relative change across two consecutive iterations is less than 10^{-5}. We use grid search on log scale to tune the learning rate and regularization coefficient, maximizing the likelihood on a validation set.[3]

We also explored the semi-supervised technique known as *self-training* to improve classification accuracy for this task. Self-training is a wrapper method that first uses labeled instances to train the model, then applies that model to a larger sample of unlabeled instances in order to impute labels. The model is then retrained on the combined set of human-labeled and machine-labeled instances.

[2] We used a custom English stopword list of 363 words.

[3] Due to the small number of labeled instances available, the validation set consists of one positive and one negative instance.

Learning Method	Learning Rate	L_2 Regularizer Coefficient	Zero Threshold	One Threshold	Training Set Size	Feature Count	Acc. (%)	F_1 (%)
supervised	0.001	0	N/A	N/A	19	55K	0.84	0.77
supervised	0.01	0.01	N/A	N/A	19	55K	0.84	0.77
self-training	0.001	0	0.01	0.8	289	813K	0.79	0.67
self-training	0.01	0.01	0.001	0.7	271	736K	0.84	0.77
self-training	0.01	0.01	0.001	0.99	175	507K	0.84	0.72

Table 1: Results of self-training.

3 Evaluation and Results

For a given concept, we employ a leave-one-out estimator of the quality of the models trained in the supervised setting. More concretely, if we have N instances of documents labeled indicating whether or not a certain concept is present, we train N models, each using $N-1$ of the labeled instances, holding out one for testing. We then calculate the accuracy of the model on the held-out instance. This accuracy is averaged across the N splits of the data.

We report accuracy and F_1 score at different decision boundary thresholds (trading precision for recall), and compare to a baseline that labels every document with the most common value in the training data.

We consider the following two concepts: *transparency on law enforcement requests* and *a user's right to terminate their account*.

3.1 Transparency on Law Enforcement Requests

In this task, we attempt to automatically determine whether a privacy policy is considered clear (by humans) about the procedures that law enforcement officials must follow when they seek access to sensitive user data. We collected 19 privacy policies annotated for this concept: 12 marked *not transparent* and 7 marked *transparent*. The number of unique features extracted from these policies (i.e., dimension of x_d) totalled about 55K features.

Figure 1 shows our performance on this task, using the leave-one-out estimator and varying the decision boundary. For example, a decision boundary of 0.25 assigns label 1 (*transparent*) to an instance iff $p(label = 1|model) > 0.25$. The results in the table were obtained by the unregularized log-loss with a learning rate of 0.001.[4] The value of the F_1-metric peaks at 0.76 and accuracy peaks at 0.84. This constitutes a 20% improvement over the accuracy of a baseline which always chooses the *not transparent* label, the most frequent class in the training data. Table 1 summarizes the results for self training, and shows that it does not improve classification accuracy for this task.

Table 2 shows the features with the greatest-magnitude positive and negative impact scores, along with their average impact scores and the percentages of labeled instances of each class in which the feature has a non-zero value. We use an L_2-regularizer for results in this table to avoid overfitting the feature weights. Intuitively, the impact score is a measure of a feature's contribution

[4]Using L_2-regularization flattens the graph at 0.84 and 0.77 for the accuracy and F_1 score, respectively, for a wider range of decision boundary values.

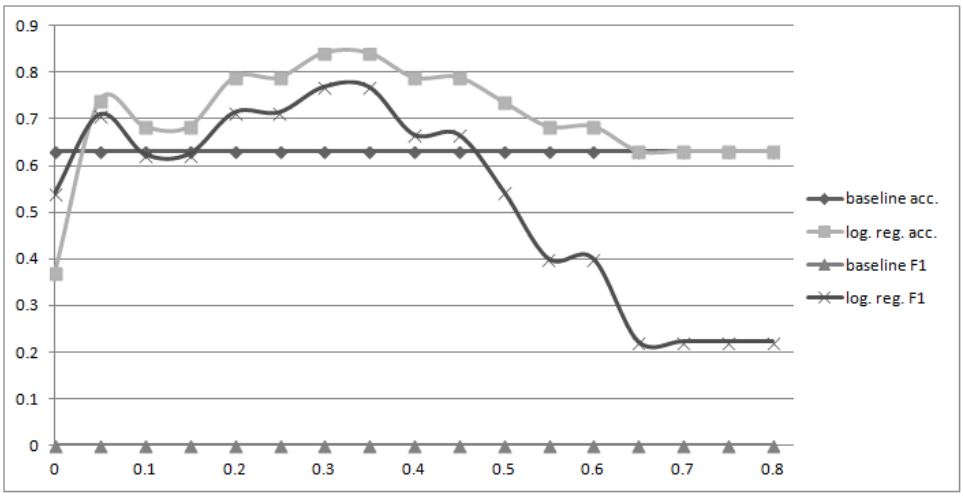

Figure 1: Performance results for the transparency task.

Feature	Average Impact	% of 0-Label	% of 1-Label	Feature	Average Impact	% of 0-Label	% of 1-Label
account	13.09	100	100	information	-19.63	100	100
access	6.04	100	100	personally	-4.86	67	43
share	4.2	100	86	identifiable	-4.63	67	43
data	2.78	92	100	personally identifiable	-4.41	67	43
using	2.07	83	86	cookies	-3.58	92	71

Table 2: Features with the greatest-magnitude positive (left) and negative (right) impact scores.

to the classification decisions for a given test set. We measure the average impact of a feature f in a train/test split of the data using the following formula:

$$\text{impact}(f) = \frac{1}{|split|} \sum_{(trainset,testset)\in splits} \text{weight}(f|trainset) \sum_{doc\in testset} \text{value}(f|testdoc) \quad (1)$$

Despite the use of L_2 regularization, some frequent but not obviously informative features such as "account" and "information" have very high positive and negative impact scores, respectively. We propose solutions for this problem in Section 4.

Table 3 shows some features which include the words *request* or *law*. Intuitively, such features might correlate with policies which are transparent regarding law enforcement requests. For all but

Feature	Average Impact	% of 0-Label	% of 1-Label	Feature	Average Impact	% of 0-Label	% of 1-Label
request	0.22	83	100	law	0.06	75	100
requests	0.04	67	57	law regulation	0.01	0	43
legal request	0.01	0	57	law requires	9.50E-04	0	29
account formal request	9.50E-04	0	29	law enforcement	-0.01	33	14

Table 3: Features including the words *request* (left) or *law* (right).

one of the features below, such features do appear more frequently in policies labeled as transparent, and some of them have a relatively large positive impact on the classification decision. Most features have an average impact of zero, which also follows our intuitions.

3.2 User's Right to Terminate Their Account

In this task, we attempt to predict whether a privacy policy gives a user the right to voluntarily cancel, terminate, or delete their account. We collected 18 negative privacy policies (i.e., those that did not express the aforementioned right) and 3 positive ones. We tried several variations in feature types[5], regularization, stopping criteria, learning rate and asymmetric loss[6], but the learned model always assigned a tiny value to $p(label = 1|testinstance)$, regardless of the class to which the test instance belonged. This concept appeared to be much more difficult to predict than *transparency on law enforcement requests*.

The difficulty of this task could be attributed to several factors. The number of positive instances may have been too small to learn adequate weights of the many features we used. One of the positive instances was a machine translation (into English) of a French privacy policy, which may have contained some translation mistakes. Furthermore, some of the annotators used external evidence to label the instances. For example, one annotator used a black list[7] of websites which do not allow users to terminate their accounts, rather than finding the information in the website's posted policies.

4 Future Work

We believe that the following are promising paths to continue investigating the problem of predicting features of privacy policies:

- *Increasing the number of annotated privacy policies available:* Currently, very few tasks can be performed automatically given the small number of annotated documents. Because our approach requires humans to produce "gold standard" annotations, it will be desirable to reduce the time and effort spent on annotating each policy. We can help annotators find the relevant pieces of a policy using models trained with readily available labeled instances. For example, we found it useful to highlight the terms with a high impact score according to the baseline model, or sections of a document containing a relatively high density of such terms.

- *Utilizing the structure of privacy policies in the learning process:* The concepts we attempted to predict in privacy policies were usually described (if at all) in a small part of the document.

[5]Given the small number of labeled instances, binary-valued features that encoded term presence in a document (as opposed to term count) were hypothesized as a reasonable representation of a document, but results showed no significant difference between using one feature type or the other.

[6]We attempted to scale the loss function for instances of one class in proportion to the number of labeled instances in the other class.

[7]http://www.accountkiller.com/en/Blacklist

Our approach of using all the lexical terms in the document therefore introduces a considerable amount of noise. In order to account for this noise, we might first use an unsupervised topic model such as latent Dirichlet allocation [1] to generate topic-features for terms in the policy document, then use those features to describe the documents in the logistic regression model. Another type of feature that can be used is section information (e.g., the texts of section titles), since most policies have standard sections for specifying what information is being collected from a user and how it is used. We can use unsupervised document segmentation techniques (e.g., Eisenstein and Barzilay's Bayesian unsupervised topic segmentation [3]) to find corresponding sections across documents and encode this information as features in the logistic regression model.

- *Pruning the available features in the bag of words and n-grams:* Feature analysis showed that many features representing frequently-occurring words and phrases had a high impact but did not correlate with either class more than the other. Feature selection methods such as group lasso regularizers and pointwise mutual information could potentially eliminate the impact of such non-informative features.

Beyond the prediction of properties of privacy policies, challenges remain in the effective presentation of this information to the user. Prior work by Kelley et al. [4] showed how internet users' privacy concerns could be addressed visually with privacy nutrition labels, structured with inspiration from P3P. This "top-down" approach is natural when the websites that the user visits provide explicit and highly structured privacy policies. However, we discovered during annotation for the present effort that natural language policies, unlike their P3P counterparts, they are often ambiguous and incomplete. This suggests a different approach—structured partly by pre-determined questions about policies (as in nutrition labels) and partly by each policy's contents—may be more appropriate. Another possible approach, if the performance of automated analysis of policies remains persistently low, may be a *semi-automated* analysis: enlisting a human to read a sentence or paragraph that automated methods have determined is likely to answer a given question about the policy. Addressing these challenges in processing privacy policy text and presenting key details to the user will advance our goal of effectively informing internet users on how their personal information is being used and collected.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *Journal of Telecommunications and High Technology Law*, 10(2), 2012.

[3] Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *EMNLP*, pages 334–343. ACL, 2008.

[4] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *SOUPS*, 2009.

[5] Pedro Giovanni Leon, Lorrie Faith Cranor, Aleecia M. McDonald, and Robert McGuire. Token attempt: the misrepresentation of website privacy policies through the misuse of p3p compact policy tokens. In *WPES*, pages 93–104, 2010.

[6] Aleecia M. McDonald and Lorrie F. Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3), 2008.