Identifying the Provision of Choices in Privacy Policy Text

Kanthashree Mysore Sathyendra
School of Computer Science
Carnegie Mellon University
ksathyen@andrew.cmu.edu

Shomir Wilson
EECS Department
University of Cincinnati
shomir.wilson@uc.edu

Florian Schaub
School of Information
University of Michigan
fschaub@umich.edu

Sebastian Zimmeck
School of Computer Science
Carnegie Mellon University
szimmeck@andrew.cmu.edu

Norman Sadeh
School of Computer Science
Carnegie Mellon University
sadeh@cs.cmu.edu

Abstract

Websites' and mobile apps' privacy policies, written in natural language, tend to be long and difficult to understand. Information privacy revolves around the fundamental principle of *notice and choice*, namely the idea that users should be able to make informed decisions about what information about them can be collected and how it can be used. Internet users want control over their privacy, but their choices are often hidden in long and convoluted privacy policy documents. Moreover, little (if any) prior work has been done to detect the provision of choices in text. We address this challenge of enabling user choice by automatically identifying and extracting pertinent choice language in privacy policies. In particular, we present a two-stage architecture of classification models to identify opt-out choices in privacy policy text, labelling common varieties of choices with a mean F1 score of 0.735. Our techniques enable the creation of systems to help Internet users to learn about their choices, thereby effectuating notice and choice and improving Internet privacy.

1 Introduction

Website privacy policies are long, verbose documents that are often difficult to understand. It has been shown that an average Internet user would require an impractical amount of time to read the privacy policies of online services that they use and would not properly understand them (McDonald and Cranor, 2008). Although Internet users are concerned about their privacy and would like to be informed about the privacy controls they can exercise, they are not willing or able to find these choices in policy text. Choices for privacy controls, which are the most actionable pieces of information in these documents, are frequently "hidden in plain sight" among other information. However, the nature of the text and the vocabulary used to present choices provide us with an opportunity to automatically identify choices, a goal that we focus upon in this paper.

We define a *choice instance* as a statement in a privacy policy that indicates that the user has discretion over aspects of their privacy. An example (which notably features a hyperlink) is the following:

> If you would like more information on how to opt out of information collection practices, go to `www.aboutads.info`.[1]

Some examples of choices offered to users include opt-outs or controls for the sharing of personal information with third parties, receiving targeted ads, or receiving promotional emails. Analyzing these choice instances in aggregate will help to understand how notice and choice is implemented in practice, which is of interest to legal scholars, policy makers and regulators. Furthermore, extracted choice options can be presented to users in more concise and usable notice formats (Schaub et al., 2015), such as a browser plug-in or a privacy based question answering system.

For this paper, we treat the identification of choice instances as a binary classification problem, in which we label each sentence in the privacy policy text as containing a choice instance or not. We use the OPP-115 Corpus (Wilson et al., 2016) for training and evaluation of our models.

[1] `http://www.nurse.com/privacy/` (last updated on July 13, 2015)

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2757–2762
Copenhagen, Denmark, September 7–11, 2017.

We further annotate a second dataset[2] and develop a composite model architecture to automatically identify and label different types of opt-out choices offered in privacy policies. We primarily focus on extracting opt-out instances with hyperlinks because these are some the most common and useful choices described in privacy policies. Moreover, these choice expressions are actionable: the first step of the action to be taken (i.e., following a hyperlink) is clearly represented in the text of these instances.

The work presented in this paper has been conducted in the context of the 'Usable Privacy Policy' project, which combines crowdsourcing, machine learning and natural language processing to overcome the limitations of today's approach to 'notice and choice' in privacy (Sadeh et al., 2013).

2 Related Work

The Federal Trade Commission identifies "Notice and Choice" as one of the core principles of information privacy protection under the Fair Information Practice Principles (Federal Trade Commission, 2000). However, privacy policies, being long, complicated documents full of legal jargon, are sub-optimal for communicating information to individuals (Cranor, 2012; Cate, 2010; Schaub et al., 2015; Reidenberg et al., 2015). Antón et al. (2002) conducted a study in which they identified multiple privacy-related goals in accordance with Fair Information Practices, which included 'Choice/Consent' as one of the protection goals.

The potential for the application of NLP and information retrieval techniques to legal documents has been recognized by law practitioners (Mahler, 2015), with multiple efforts applying NLP techniques to legal documents. Bach et al. (2013) use a multi-layer sequence learning model and integer linear programming to learn logical structures of paragraphs in legal articles. Galgani et al. (2012) present a hybrid approach to summarization of legal documents, based on creating rules to combine different types of statistical information about text. Early work on automatically extracting annotations from privacy policies includes that of Ammar et al. (2012). Montemagni et al. (2010) investigate the peculiarities of the language in legal text with respect to that in ordinary text by applying shallow parsing. Ramanath et al. (2014) introduce an unsupervised model for the automatic alignment of privacy policies and show that Hidden Markov Models are more effective than clustering and topic models. Liu et al. (2016a) modelled the language of vagueness in privacy policies using deep neural networks.

Many of these efforts consider legal documents as a whole, and they focus less on identifying specific attributes of data practices such as choices. We focus on choices in the present work because of their potential to present Internet users with engaging, directly actionable information.

3 Approach

We used the OPP-115 Corpus to train and evaluate our models for identifying opt-out choices. The corpus consists of 115 website privacy policies and annotations (created by law students) for *data practices* that appear in them. A data practice is a statement about how a website user's personal information is collected, processed or shared. Each data practice consists of a selection of a category (i.e., a theme associated with the practice, such as "First Party Collection/Use"), a set of values for attributes specific to the category, and text spans from the policy associated with the value selections (Wilson et al., 2016). The attributes representing choice instances are present in multiple categories of data practices, namely "First Party Collection/Use," "Third Party Sharing/Collection," "User Access, Edit and Deletion," "Policy Change," and "User Choice/Control." The dataset contains annotations for different types of user choice instances, namely "opt-in," "opt-out," "opt-out link," "opt-out via contacting company," "deactivate account," "delete account (full)," and "delete account (partial)."

3.1 Dataset Refinement

We treated the problem of extracting choice instances as a binary classification problem where we labeled sentences from a privacy policy as containing a choice instance (positive) or not (negative). We focused specifically on opt-out choices, as they are among the most common choices offered to Internet users and because opting out is notoriously difficult for users (Leon et al., 2012). All sentences that contained an opt-out user choice (as specified by the OPP-115 annotations) were considered positive, and the rest were considered negative. This resulted in a gold standard set of

[2] Available for download at https://www.usableprivacy.org/data

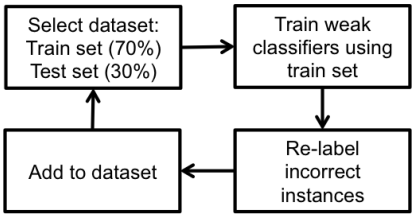

Figure 1: Active learning with relabelling.

labeled sentences with 251 positive instances and approximately 12K negative instances.

Differences between our problem formulation and the OPP-115 annotation scheme led to the need for a few label adjustments. Opt-out text spans which crossed sentence boundaries resulted in positive labels for all involved sentences, although often only one of the sentences in a span was positive. Additionally, during the OPP-115 annotation procedure, the fact that hyperlinks were not shown to annotators meant that some choice instances were not correctly identified. This resulted in noisy labels in our derived dataset.

The unbalanced distribution of the opt-out labels allowed us to manually verify and correct labels in the positive class. However, correcting errors in the much larger negative class (of 12K instances) was a challenge, since comprehensive manual verification was infeasible. Instead, we adopted a semi-automated, iterative relabelling approach with active learning. We randomly divided the dataset into train (70%) and test (30%) sets. We trained a binary logistic regression classifier using bag of n-gram features on the training data, and then used it to classify the test data. This was essentially a weak classifier, since it was trained on noisy (unverified) data. We manually examined the false positives and false negatives as given by this model and relabelled incorrectly labelled instances, thus reducing noise in the dataset. Performing multiple iterations of this approach, each time with a different train and test set, resulted in a much cleaner dataset (Figure 1). Following this refinement, the model F1 scores improved and were also more accurate. For all our experiments thereon, we used this refined version of the dataset for training and evaluation.

3.2 Coarse-Grained Classification

We divided the dataset into train and test sets of 85 and 30 privacy policies, respectively. We experimented with a variety of features for *coarse-grained classification*, to separate positive and negative instances:

Stemmed Unigrams and Bigrams. We removed most stop words from the feature set, although some were retained for the modal verb and opt-out features (described below). Bigrams are important to capture pertinent phrases such as "opt out."

Relative Location in the Document. This was a ratio between the number of sentences appearing before the sentence instance and the total number of sentences in the privacy policy.

Topic Model Features. We represented the OPP-115 segment (roughly, a paragraph) containing the sentence instance as a topic distribution vector using latent Dirichlet allocation (Blei et al., 2003) and non-negative matrix factorization (Xu et al., 2003) with 8 and 10 topics, respectively. Previous work on vocabulary intersections of expert annotations and topic models for data practices in privacy policies (Liu et al., 2016b) inspired us to take this approach.

Modal Verbs and Opt-Out specific phrases. We observed vocabulary cues in positive instances that suggested a domain-independent "vocabulary of choice". Many positive instances were imperative sentences and contained modal words such as *may*, *might*, or *can*. We also identified key phrases in the training set such as *unsubscribe* and *opt-out* that were indicative of opt-out choices.

Syntactic Parse Tree Features. We obtained constituency parse trees for sentences using the Stanford Parser (Manning et al., 2014) and extracted production rules and non-terminals as features. We included the maximum depth and average depth of the parse tree as features, as these are indications of specificity.

We used logistic regression classification for the coarse-grained classification stage. Model hyperparameters were tuned based on 5-fold cross validation on the training set. The final parameters for the best performing model had the inverse L2 regularization constant set at C=1.3 and class-weights of 1.5 and 1 for positive and negative class, respectively.

3.3 Fine-Grained Classification

We also developed a *fine-grained* model to differentiate between varieties of opt-out instances. For training data, we annotated a set of 125 positive instances to assign two additional labels to each of them; these were *Party Offering Choice* and *Purpose*. Party Offering Choice could be one of *First*

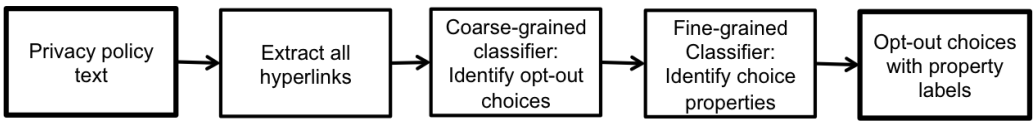

Figure 2: Two-Tier Classification Model.

Annotation	# Instances
TH,AD	52
FI,CM	19
FI,AD	15
FI,SH	6
TH,AN	4
BR,CK	2
TH,SH	2
FI,CK	1
TH,CK	1

Table 1: Distribution of different annotation types.

Party (FI), *Third Party*, (TH), or *Browser* (BR). Purpose could be one of *Advertisement* (AD), *Data Sharing* (DS), *Communications* (CM), *Analytics* (AN) or *Cookies* (CK). Table 1 shows the distribution of these annotations. To predict these labels, we trained eight binary logistic regression classifiers, one for each of the preceding values. If multiple classifiers in a label set returned positive, we selected the prediction with the higher log likelihood. The features we used for these classifiers were:

Stemmed Unigrams and Bigrams. We collected bags of n-grams from the sentence under consideration and its containing segment.

Anchor Text. The anchor text of the hyperlink in the sentence.

Hyperlink URL Tokens. We split the URL by punctuation (such as '/' and '.') and extracted tokens.

Privacy Policy URL Tokens. We also extracted tokens from the policy URL as features.

URL Similarity Measure. We calculated the Jaccard index between the vocabulary of the policy URL and the hyperlink URL. This feature is used to identify whether the hyperlink was to a first-party page or a third-party page.

Figure 2 illustrates the overall architecture of our system. We first use the coarse-grained step to identify the presence of an opt-out instance, and then use the fine-grained step to ascertain key properties of an opt-out choice if one is present.

4 Results and Discussion

This work is one of the first efforts to automatically detect the provision of choices in text. For the coarse-grained task, we consider a simple baseline that labels sentences as positive if they contain one or more opt-out specific words, which come from a vocabulary set that we identified by examining positive instances in the training set. The F1 of the baseline was 0.554.

We performed ablation tests excluding one feature at a time from the coarse-grained classifier. The results of these tests are presented in Table 2 as precision, recall, and F1 scores for the positive class, i.e., the opt-out class. Using the F1 scores as the primary evaluation metric, it appears that all features help in classification. The unigram, topic distribution, nonterminal, and modal verb and opt-out phrase features contribute the most to performance. Including all the features results in an F1 score of 0.735. Ablation test without unigram features resulted in the lowest F1 score of 0.585, and by analyzing features with higher logistic regression weights, we found n-grams such as *unsubscribe* to have intuitively high weights. We also found the production rule "S→SBAR, VP" to have a high weight, indicating that presence of subordinate clauses (SBARs) help in classification.

For an additional practical evaluation, we created a second dataset of sentences from the privacy policies of the 180 most popular websites (as determined by Alexa rankings. We selected only those sentences that contained hyperlinks, since they are associated with particularly actionable choices in privacy policy text. We used our model (as trained on the OPP-115 Corpus) to label the 3,842 sentences in this set, and then manually verified the 124 positive predictions, observing perfect precision. Although we were unable to measure recall using this method, the high precision suggests the robustness of the model and the practical applicability of this approach to tools for Internet users.

The results for the opt-out type classification are shown in Table 3. Because of data sparsity, we show performance figures for only the top two most frequent label combinations. These results

Features/Models	Precision	Recall	F1
All	**0.862**	**0.641**	**0.735**
All - Unigrams	0.731	0.487	0.585
All - Bigrams	0.885	0.590	0.708
All - Rel. Location	0.889	0.615	0.727
All - Topic Models	0.852	0.590	0.697
All - Productions	0.957	0.564	0.710
All - Nonterminals	0.913	0.538	0.677
All - Max. Depth	0.857	0.615	0.716
All - Avg. Depth	0.857	0.615	0.716
Phrase Inclusion - Baseline	0.425	0.797	0.554
Paragraph Vec. - 50 Dimensions	0.667	0.211	0.320
Paragraph Vec. - 100 Dimensions	0.667	0.158	0.255

Table 2: Results of ablation tests for the coarse-grained classifier.

	Precision	Recall	F1
FI, CM	0.947	0.947	0.947
TH, AD	0.905	0.977	0.940

Table 3: Fine-grained classifier results.

also demonstrate a practical level of performance for Internet user-oriented tools.

5 Conclusion

We presented an approach to the problem of automatically identifying privacy choices in privacy policy text. Our experiments show that a two-stage supervised learning procedure is appropriate for this task. Our approach is to initially identify choices offered by the text and then to determine their properties. Using ablation tests, we showed that a mixture of feature types can improve upon the performance of a baseline bag-of-words model. Planned future work for this project will include the creation of a browser plug-in to present opt-out hyperlinks to Internet users.

6 Acknowledgements

This work has been supported by the National Science Foundation as part of the Usable Privacy Policy Project (www.usableprivacy.org) under Grant No. CNS 13-30596. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, or the US Government

References

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study.

Annie I Antón, Julia Brande Earp, and Angela Reese. 2002. Analyzing website privacy requirements using a privacy goal taxonomy. In *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on*, pages 23–31. IEEE.

Ngo Xuan Bach, Nguyen Le Minh, Tran Thi Oanh, and Akira Shimazu. 2013. A two-phase framework for learning logical structures of paragraphs in legal articles. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(1):3.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Fred H Cate. 2010. The limits of notice and choice. *IEEE Security & Privacy*, 8(2):59–62.

Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273.

Federal Trade Commission. 2000. Privacy Online: A Report to Congress. Technical report, Federal Trade Commission.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 115–123. Association for Computational Linguistics.

Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. 2012. Why johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 589–598. ACM.

Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016a. Modeling language vagueness in privacy policies using deep neural networks. In *2016 AAAI Fall Symposium Series*.

Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2016b. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *2016 AAAI Fall Symposium Series*.

Lars Mahler. 2015. What is nlp and why should lawyers care? http://www.lawpracticetoday.org/article/nlp-lawyers/.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Aleecia M McDonald and Lorrie Faith Cranor. 2008. Cost of reading privacy policies, the. *ISJLP*, 4:543.

Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. *Semantic Processing of Legal Texts*. Springer.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A Smith. 2014. Unsupervised alignment of privacy policies using hidden markov models.

Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, Rohan Ramanath, Cameron Russell, Norman Sadeh, and Florian Schaub. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39.

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy policy project. Technical report, Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa. USENIX Association.

S Wilson, F Schaub, A Dara, F Liu, S Cherivirala, P G Leon, M S Andersen, S Zimmeck, K Sathyendra, N C Russell, T B Norton, E Hovy, J R Reidenberg, and N Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics, Aug 2016*. ACL.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM.