Automatic Extraction of Opt-Out Choices from Privacy Policies

Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, Norman Sadeh
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
ksathyen@cs.cmu.edu, fschaub@cs.cmu.edu, shomir@cs.cmu.edu, sadeh@cs.cmu.edu

Abstract

Online "notice and choice" is an essential concept in the US FTC's Fair Information Practice Principles. Privacy laws based on these principles include requirements for providing notice about data practices and allowing individuals to exercise control over those practices. Internet users need control over privacy, but their options are hidden in long privacy policies which are cumbersome to read and understand. In this paper, we describe several approaches to automatically extract choice instances from privacy policy documents using natural language processing and machine learning techniques. We define a *choice instance* as a statement in a privacy policy that indicates the user has discretion over the collection, use, sharing, or retention of their data. We describe machine learning approaches for extracting instances containing opt-out hyperlinks and evaluate the proposed methods using the OPP-115 Corpus, a dataset of annotated privacy policies. Extracting information about privacy choices and controls enables the development of concise and usable interfaces to help Internet users better understand the choices offered by online services.

1 Introduction

Website privacy policies are long, verbose documents which are often difficult to understand. It has been shown that an average Internet user would require a substantial amount of time to read the privacy policies of online services they use (McDonald and Cranor 2008) and might not even completely understand them. Although users are concerned about their privacy online and would like to be informed about various privacy controls they can exercise, they are not willing to read lengthy privacy policy documents due to the amount of time and effort involved in reading and understanding them. Options for privacy controls are hidden in long text in privacy policy documents. However, the nature of the text and vocabulary surrounding the text indicating such options provide us with an opportunity to apply machine learning techniques and natural language processing techniques to automatically identify user choices and controls described in privacy policy documents. Thus, we focus on the problem of automatically extracting user choice instances from privacy policy documents.

We define a *choice instance* as a statement in a privacy policy that indicates that the user has discretion over certain aspects related to their privacy. Some examples of choices offered to users include opt-outs or controls over some types of sharing of user's personal information with third parties, receiving targeted ads, or receiving promotional emails. Analyzing these choice instances from policies helps 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).

We treat this as a problem in which we classify all the sentences in privacy policy text as containing a choice instance or not. We propose to use supervised machine learning methods along with phrase inclusion models to automatically identify choice instances in privacy policies. Phrase inclusion models classify instances chiefly based on the presence of certain phrases in them. We use the OPP-115 dataset (Wilson et al. 2016), which contains fine-grained data practice annotations for 115 privacy policies for training and evaluation of our machine learning models.

We also identify vocabulary surrounding choice instances and comment on language specific to presence of choices in text. In particular, we leverage the presence of modal verbs in sentences containing choice instances to build phrase inclusion models to identify such instances. We further combine these phrase inclusion models with machine learning models to improve the precision of the results.

The rest of the paper is organised as follows. Section 2 describes prior work related to the application of natural language processing (NLP) to better understand legal documents. In Section 3 we present the proposed approaches including the data preprocessing steps, phrase inclusion models, and machine learning models. The results are discussed in Section 4. We present our conclusions and discuss directions for future work in Section 5.

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 not an optimal mechanism for communicating such information to individuals (Cranor 2012; Cate 2010; Schaub et al. 2015). Antón, Earp, and Reese (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. Their work identified certain key words such as 'opt-in' and 'opt-out' to be indicative of choice and consent goals in privacy documents.

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 multi-layer sequence learning model and integer linear programming to learn logical structures of paragraphs in legal articles. Galgani, Compton, and Hoffmann (2012) present a hybrid approach to summarization of legal documents, based on creating rules to combine different types of statistical informtion about text. Montemagni, Peters, and Tiscornia (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. Cranor et al. (2013) leveraged the standardized format of privacy notices in the U.S. financial industry to automatically analyze privacy polices of financial institutions. Supervised learning methods and rule based learning methods were also proposed to extract some of a websites data practices from its privacy policy (Costante et al. 2012; Zimmeck and Bellovin 2014).

However, many of these efforts consider legal documents as a whole and focus less on identifying specific attributes of data practices such as choices. These previous works indicate the potential of automatically identifying user choices in privacy policies and legal documents.

3 Approach

We propose to use supervised machine learning methods to automatically extract choice instances from privacy policy text. We used the OPP-115 dataset to train and evaluate our models. The dataset consists of 115 website privacy policies and annotations for 10 different data practice categories, where each data practice is articulated by a category-specific set of attributes (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/Use,' '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).

In this paper, we focus on extracting opt-out instances containing hyperlinks (henceforth referred to as 'opt-out' instances) from privacy policy text. We focus on opt-out because we believe that users would find opt-out hyperlinks more useful than other types of choice instances. Opt-out

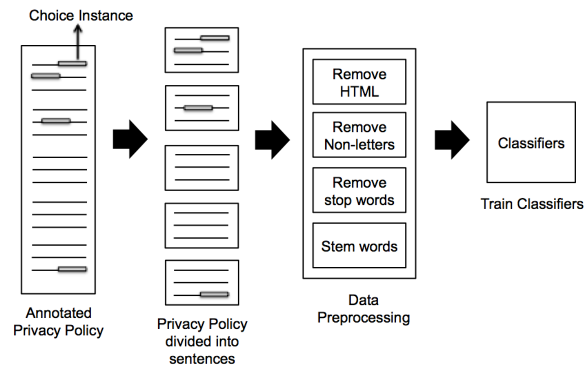

Figure 1: Overview diagram for the proposed method.

instances are also one of the most common choice types described in privacy policies.

We treat the problem of extracting choice instances as a binary classification problem where we classify sentences from the privacy policy as containing a specific kind of choice instance (positive) or not (negative). The privacy policy is first divided into segments, which roughly correspond to paragraphs. These segments are further divided into sentences. We then work with these sentences as our instances for classification. We divide the dataset into training and test sets with training data consisting of sentences from 75 policies and test data consisting of sentences from the remaining 40 privacy policies. The steps involved in the classification process are described in the following subsections.

Data Preprocessing

The dataset consists of 3,792 paragraph segments from 115 policies. For our experiments, these segments were further divided into sentences using the Natural Language Toolkit sentence tokenizer.[1]. Since these sentences contained HTML tags, which would not help in classification, we used BeautifulSoup[2] to remove them. Non-letter characters (incl. punctuation and special characters) were also removed, as well as English stop words. Snowball stemmer was then used to stem the resulting words. Stemming reduces words to their morphological root (Biagioli et al. 2005) in order to cluster words with common semantics (Francesconi and Passerini 2007). The preprocessed sentences were then used to extract features to train models.

Classification

We experiment with two different types of classification models – phrase inclusion models and machine learning models based on the idea that vocabulary used to describe opt-out choices is fairly constrained.

Phrase Inclusion Models We manually identified some phrases such as 'opt-out', 'unsubscribe' and 'opting-out' that were specific to opt-out choice instances, by carefully

[1] `http://www.nltk.org/index.html`

[2] `https://www.crummy.com/software/BeautifulSoup/`

examining the training sentences from the corpus. Some examples using such phrases are as follows:

> If you no longer wish to receive a specific newsletter, you may opt out by following the **'unsubscribe'** instructions located near the bottom of each newsletter.

> If you would prefer us not to contact you for these purposes, simply go to our **opt-out** page and let us know.

As a baseline approach, we developed a simple phrase inclusion model, to classify sentences based on the presence of these phrases. Given a sentence, we classify it as positive if it contains any of these phrases and negative otherwise. This simple model is based on the intuition that opt-out choice instances are expressed using a specific and limited set of words and identifying these words or phrases gives us information about the instance being an opt-out instance.

Machine Learning Models We trained several machine learning models to classify sentences. The models included logistic regression, linear support vector machines, random forest, naïve Bayes and Nearest Neighbor models. To train these models, we used unigram and bigram bag-of-words features. In the bag-of-n-grams model, we break each sentence into n-grams and form a vector for each sentence representing these n-grams. These bag-of-n-grams models work due to the fairly constrained vocabulary surrounding opt-out instances. Bigrams are important to capture information represented by two-word phrases, like 'opt out,' which are not captured by unigram models alone. Given the small size of the corpus, unigrams and bigrams, capture sufficient information for classification with fewer features. Initially, we ran experiments with the feature set consisting of unigram features only. We then conducted experiments after adding bigram features to the feature set. We did not add trigram features due to high memory and computational requirements.

We made an interesting observation about the tone of the sentences depicting positive instances. The positive instances were mostly imperative sentences and most sentences contained modal verbs like 'may', 'might', or 'can' that indicated choices to users. A modal verb is a type of verb that is used to indicate modality – that is: likelihood, ability, permission, and obligation. An example of a positive instance containing modal verbs is:

> You **may** opt out of receiving these general communications by using one of the following methods: Select the email opt out or unsubscribe link, or follow the opt-out instructions included in each email communication.

To incorporate this information provided by modal verbs and opt-out specific vocabulary, we further expanded the feature space by adding a custom feature representing the presence of opt-out specific vocabulary and modal verbs. We bundled together phrases characteristic of opt-out instances into one feature and used this custom feature in addition to the existing unigram and bigram bag-of-words features. We used the NLTK Part-Of-Speech(POS) tagger to obtain the POS tags of words in the sentences and gave the feature a value of one if any of the words in the sentence was tagged as a modal verb and zero otherwise.

While the above set of features gave us better F1 scores, we were focusing on improving the precision of the results. If our ultimate goal were to represent the information most useful to users, we would want all the displayed information to be accurate. Thus, we emphasized precision over recall. To improve precision of the results, we ran two classification models in series and classified sentences as positive only if both classifiers did so. Since the phrase inclusion models and Logistic Regression models performed well, we used these two models in series. The results for all these methods are presented in the following section.

4 Results and Discussion

Table 1 shows the precision, recall and F1 score for the positive class for phrase inclusion models. These models are simple methods used to classify based on presence of opt-out specific words. The high recalls for these methods indicate that a majority of the positive instances contain these opt-out specific words. The lower precision indicates that these are present in a few negative instances as well.

The results for different machine learning models using different feature sets are shown in Table 2. The Logistic Regression model performed best with a precision of 0.59, recall of 0.507 and F1 score of 0.530 for the positive class. This was closely followed by the SVM model. We also observe that the F1 scores for unigram+bigram features are higher, indicating that the bigram features capture additional information about the sentences.

The results for the logistic regression model using unigram+bigram features along with the custom feature representing modal verbs and opt-out specific vocabulary is also displayed in Table 2. As seen in the table, this score is a significant improvement over the scores for models without the custom feature. This is indicative of the fact that there exists a high correlation between the custom feature and positive instances. In other words, presence of the opt-out specific words or modal verbs indicate that the instance is more likely to be positive.

The results for the combination models are presented in the last row of Table 2. As seen in the table, the combination of a Phrase Inclusion model with Logistic Regression has the best precision score of 0.692 and the best F1 score of 0.585. The phrase inclusion model with its high recall selects the most probable positive instances containing opt-out specific phrases. Further, the Logistic Regression model filters out possible negative instances, thus improving the precision. A higher precision combined with a good recall results in a better F1 score. The combination model resulted in a decrease in the number of false positives by 6 instances while the number of true positives reduced by 1 instance. This resulted in a significant improvement in precision.

We analyzed false positive instances produced by these methods. Consider the following two examples:

1. You can obtain more information about these advertising service providers' information collection practices, and opt out of such practices (and at the same time opt out of the collection practices of other, or all, NAI members) by following the opt out instructions on the NAI's web-

Model	Opt-Out specific Vocabulary used for classification	Precision	Recall	F1	Accuracy
Phrase Inclusion Model 1	**opt-out, opt out, unsubscribe, opting out**	**0.425**	**0.797**	**0.554**	**0.977**
Phrase Inclusion Model 2	opt-out, opt out, unsubscribe, opting out, click here	0.383	0.825	0.523	0.974
Phrase Inclusion Model 3	opt-out, opt out, unsubscribe, opting out, if you do not want, click here	0.363	0.841	0.507	0.971
Phrase Inclusion Model 4	opt-out, opt out, unsubscribe, opting out, if you do not want	0.398	0.813	0.534	0.975

Table 1: Results for Phrase Inclusion Models. The best result is highlighted in bold.

Feature Set	Model	Precision	Recall	F1	Accuracy
Unigram	**Logistic Regression**	**0.574**	**0.493**	**0.530**	**0.987**
	SVM	0.417	0.493	0.452	0.982
	Nave Bayes	0.263	0.634	0.372	0.967
	Ranfom Forest	0.667	0.254	0.367	0.987
Unigram + bigram Bag of words	**Logistic Regression**	**0.59**	**0.507**	**0.545**	**0.987**
	SVM	0.537	0.507	0.522	0.986
	1 Nearest Neighbor	0.542	0.451	0.492	0.986
	4-NN with 1000 features	0.581	0.352	0.439	0.986
	Nave Bayes	0.324	0.662	0.435	0.974
	4 NN	0.571	0.338	0.425	0.986
	Random Forest	0.645	0.282	0.392	0.987
	5 NN	0.543	0.268	0.358	0.985
Custom Feature: Unigram and Bigram bag of words + Modal Verbs and opt-out specific phrases	**Logistic Regression**	**0.632**	**0.507**	**0.563**	**0.988**
Custom Feature and Phrase Inclusion Model 1	**Combination Model: Logistic Regression and Phrase Inclusion Model 1**	**0.692**	**0.507**	**0.585**	**0.989**

Table 2: Results for various Machine Learning Models. The best results in each category are highlighted in bold

site at `http://www.networkadvertising.org/managing/opt_out.asp`. If you would like more information on how to opt out of information collection practices, go to `www.aboutads.info`.

2. For more information about DoubleClick, cookies, and how to opt-out, please click here.

These examples were classified as positive instances by our classifiers but were not identified as positive instances in the labeled data. A reason may be that hyperlinks were not visible to the annotators during the annotation procedure, which may have resulted in few false positives and false negatives. Absence of such false positives and negatives in the labeled data could result in better precision and F1 scores. Even with the limited number of annotations in the dataset, the models were able to produce meaningful results.

5 Conclusion

We considered the task of automatically extracting user choice instances from privacy policy text. We approached this as a classification problem and used phrase inclusion models and supervised machine learning approaches to accomplish this task. Our experiments showed that machine learning is feasible for this task, even with a dataset containing a limited number of annotations. Further, we identified vocabulary specific to opt-out instances and used these as features for our machine learning models. We showed that by using verb modality and the vocabulary specific to opt-out instances, we can obtain better results in comparison with a feature set comprised only of bag-of-words models. Our experiments also showed that running two classifiers in series increases accuracy for this problem. As part of future work, we plan to incorporate semantic features into our models, as well as study the importance of the sentence structure in identifying such instances. Lack of hyperlinks in the annotation task caused the dataset to contain a few false positives and false negatives. Addressing these aspects may further increase the accuracy of our methods.

References

Antón, A. I.; Earp, J. B.; and Reese, A. 2002. Analyzing website privacy requirements using a privacy goal taxonomy. In *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on*, 23–31. IEEE.

Bach, N. X.; Minh, N. L.; Oanh, T. T.; and Shimazu, A. 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.

Biagioli, C.; Francesconi, E.; Passerini, A.; Montemagni, S.; and Soria, C. 2005. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law*, 133–140. ACM.

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

Costante, E.; Sun, Y.; Petković, M.; and den Hartog, J. 2012. A machine learning solution to assess privacy policy completeness:(short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, 91–96. ACM.

Cranor, L. F.; Idouchi, K.; Leon, P. G.; Sleeper, M.; and Ur, B. 2013. Are they actually any different? comparing thousands of financial institutions privacy practices. In *The Twelfth Workshop on the Economics of Information Security (WEIS 2013)*.

Cranor, L. F. 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.

Francesconi, E., and Passerini, A. 2007. Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law* 15(1):1–17.

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

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

McDonald, A. M., and Cranor, L. F. 2008. Cost of reading privacy policies, the. *ISJLP* 4:543.

Montemagni, S.; Peters, W.; and Tiscornia, D. 2010. *Semantic Processing of Legal Texts*. Springer.

Ramanath, R.; Liu, F.; Sadeh, N.; and Smith, N. A. 2014. Unsupervised alignment of privacy policies using hidden markov models.

Schaub, F.; Balebako, R.; Durity, A. L.; and Cranor, L. F. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, 1–17. Ottawa: USENIX Association.

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

Zimmeck, S., and Bellovin, S. M. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, 1–16.