Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies

Frederick Liu, Shomir Wilson, Florian Schaub, Norman Sadeh
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
{fliu1, shomir, fschaub, sadeh}@cs.cmu.edu

Abstract

Privacy policies are commonly used to inform users about the data collection and use practices of websites, mobile apps, and other products and services. However, the average Internet user struggles to understand the contents of these documents and generally does not read them. Natural language and machine learning techniques offer the promise of automatically extracting relevant statements from privacy policies to help generate succinct summaries, but current techniques require large amounts of annotated data. The highest quality annotations require law experts, but their efforts do not scale efficiently. In this paper, we present results on bridging the gap between privacy practice categories defined by law experts with topics learned from Non-negative Matrix Factorization (NMF). To do this, we investigate the intersections between vocabulary sets identified as most significant for each category, using a logistic regression model, and vocabulary sets identified by topic modeling. The intersections exhibit strong matches between some categories and topics, although other categories have weaker affinities with topics. Our results show a path forward for applying unsupervised methods to the determination of data practice categories in privacy policy text.

Introduction

Privacy policies are used to inform Internet users about the data and privacy practices of websites and online services they visit. Most countries require that website operators post a notice of how they gather and process users' information, resulting in a very large number of privacy policy documents, if we were to crawl the whole web. On the other hand, as human annotation tends to be a rather time consuming process requiring that multiple people look at the same text to produce reliable annotations (Reidenberg et al. 2015), (Wilson et al. 2016b). The Usable Privacy Policy Project (Sadeh et al. 2013) introduced a corpus of 115 website privacy policies annotated with detailed information about the data practices that they describe (Wilson et al. 2016a). The annotations in this corpus reveal the structure and complexity of these documents. Their annotators required a mean of 72 minutes to annotate a privacy policy, resulting in prohibitive costs for scaling manual annotations. The scheme for annotating the structure and contents of privacy policies was defined through analysis and iterative refinement by domain experts. The annotations were then created by law students.

In this work, our goal is to leverage the unlabeled data into our automated annotation process along with the expensive annotated data. Unsupervised methods such as topic models provide us with an approach to learn the structure of privacy policies without any annotations. However, we do not know the relationship of the learned structure with our defined labels. To understand the relationship, we calculate the vocabulary-based similarity of topics from topic modeling with expert-defined categories of data practices. We visualize the relationship with a color map and find promising results that could be further compared and evaluated with current supervised methods in the future.

The paper is structured as follows. We discuss the two machine learning techniques – one supervised and one unsupervised – that we use to extract the vocabularies from these policies. We then describe how we calculate the similarity of categories and topics using the intersections of their vocabularies. Finally, we analyze and discuss the results of our experiments, which show substantial evidence of mappings between certain categories and topics.

Related Work

The research reported herein was conducted as part of the Usable Privacy Policy Project, a multi-year, multi-organizational effort to semi-automatically annotate privacy policies at scale through the exploration of techniques that combine crowdsourcing, machine learning and natural language processing (Sadeh et al. 2013; Wilson et al. 2016c; Breaux and Schaub 2014; Schaub, Breaux, and Sadeh 2016). Because of the number of privacy policies that need to be annotated and because many of these policies are updated on a regular or semi-regular basis, we aim to increasingly automate the annotation process. Some prior knowledge has leveraged natural language processing and machine learning on privacy polices. Some work has leveraged natural language processing in the pipeline of information extraction techniques, to extract the list of data collected by a website, according to what it is stated in its privacy policy (Costante,

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

den Hartog, and Petković 2013). Some have tried extractioning answers for categorical questions about privacy policies (Ammar et al. 2012; Zimmeck and Bellovin 2014) with natural language processing. On the machine learning side, some have tried annotating issues in privacy policy segments by approaching the problem as an alignment problem with Hidden Markov Models (Ramanath et al. 2014) or have tried automating the categories of privacy policies with supervised methods. Other approaches leveraged Latent Dirichlet allocation (Chundi and Subramaniam 2014) to facilitate access and comprehension of privacy policies of some of the most popular websites. There is also work on closing the gap between academia and ready-to-use software packages with Vector Space Models (Řehůřek and Sojka 2010).

Compared to previous work, we formulate the analysis of privacy policies as a classic machine learning problem – multilabel classification; e.g. a segment in privacy policy can contain information of multiple categories such as First Party Collection/Use and Third Party Sharing/Collection. We tackle this problem with supervised methods and leveraging unlabeled data with an unsupervised method. Our work sheds light on solving the problem of limited annotated data by involving unlabeled data in the automatic annotation process.

Approach

In this section, we introduce the OPP-115 Corpus and explain how we extract the labels for each category from the corpus. We then explain our approach in leveraging unlabeled data to our goal of labeling policy segments by focusing on the vocabulary used in each segment.

The OPP-115 Corpus

In our experiments we leverage annotation data from the OPP-115 Corpus (Wilson et al. 2016a), which is a set of 115 online privacy policies extensively annotated by law students. Each privacy policy is annotated by three law students and the corpus consists of manual annotations for 23K fine-grained *data practices*. Privacy policies were divided into paragraph-length *segments* for annotators to read, one at a time in sequence. For each *segment*, an annotator may label zero or more *data practices* from each *category*. An individual *data practice* belongs to one of ten *categories*, and it is articulated by a category-specific set of *attributes*. For example, a User Choice/Control *data practice* is associated with four mandatory *attributes* (Choice Type, Choice Scope, Personal Information Type, Purpose) and one optional *attribute* (User Type). In the following experiments, we focus on on what *categories* of *data practices* have been annotated in a policy segment, i.e., we leverage the annotations' *data practice categories* as labels for a segment, as could be obtained from a less detailed annotation task. Since each *segment* is annotated multiple *data practices* from each *category*, each segement can have multiple *categories* as labels. We aggregated the annotations of the three annotators by setting the category label as positive when two or more annotators agree that the given policy segment includes information of the *category*. We further separate out the different *attribute values* of the Other category as individual *categories*. As a result, we consider the following twelve data practice category labels in our work:

- *First Party Collection/Use*: how and why a service provider collects user information.
- *Third Party Sharing/Collection*: how user information may be shared with or collected by third parties.
- *User Choice/Control*: choices and control options available to users.
- *User Access, Edit, & Deletion*: if and how users may access, edit, or delete their information.
- *Data Retention*: how long user information is stored.
- *Data Security*: how user information is protected.
- *Policy Change*: if and how users will be informed about changes to the privacy policy.
- *Do Not Track*: if and how Do Not Track signals[1] for online tracking and advertising are honored.
- *International & Specific Audiences*: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
- *Introductory/Generic* (subcategory of Other): introduces a the policy or makes generic statments.
- *Practice Not Covered* (subcategory of Other): describes a data practice which is not covered by all the data practice categories.
- *Privacy Contact Information* (subcategory of Other): how to contact the company.

Despite the fact that all 12 cateogry labels are related to privacy practices, we proceeded with the assumption that each category has its own vocabulary that explains its content. In the following sections, we explain how we leverage unlabeled data with topic models and analyze the relationship between the listed categories and the topics.

Vocabularies from Supervised Method

With the annotated data from the OPP-115 corpus, we are able to use supervised machine learning methods to predict the category labels of each annotated policy segment. We represent each policy segment with a TF-IDF vector and build a binary classifier for each category to classify the segments of a privacy policy to particular practice categories. We used logistic regression with an L1 regularizer to analyse the vocabulary of each category. The coefficients of the linear combination of the features can be regarded as the characteristic of each category. For example, the top 10 words for each category can be seen in Table 1.

Vocabularies from Unsupervised Method

We further evaluate the performance of unsupervised machine learning techniques that allow us to analyze policy segments without the help of expert annotations. We experimented with both Latent Dirichlet allocation and Non-Negative Matrix Factorization. NMF generated topics which

[1] www.w3.org/2011/tracking-protection

resemble our categories while LDA generated topics for different domains of the privacy policy. The following sections first briefly introduce NMF, and illustrate how we calculate the closness of categories and topics.

Non-Negative Matrix Factorization

$$\mathbf{V} = \mathbf{WH} \quad (1)$$

We use non-negative matrix factorization (NMF) to extract the topics in the set of privacy policies. We represent each policy segment as a numeric vector where each entry of the vector is the Term Frequency Inverse Document Frequency (TF-IDF) of a vocabulary in our corpus. The whole training corpus is represented with a matrix $\mathbf{V}$, where V_{ij} is the TF-IDF of the jth vocabulary in the dictionary that appears in the ith segment of our corpus. The matrix $\mathbf{V}$ is then factorized into two matrices $\mathbf{W}$ and $\mathbf{H}$ as in Equation 1, with the property that all three matrices have no negative elements. The column vectors of $\mathbf{V}$ can be represented as linear combinations of the column vectors in $\mathbf{W}$ using coefficients supplied by columns of $\mathbf{H}$. Here, we treat the column vectors of $\mathbf{W}$ as a distribution over the vocabulary of one of the topics. The size of $\mathbf{W}$ and $\mathbf{H}$ is decided by specifying a hyperparameter K, which is the number of topics. Figure 1 displays the top 10 words with highest weight from the NMF model for K=12. By looking at the table, we can observe that Topic 3 is related to Specific Audiences of privacy policies, which is one of the categories in our labels. We will discuss this in more detail in the following sections.

Bridging the Two Worlds

Both the supervised method and the unsupervised methods provide us a numeric vector of size $|\mathbf{V}|$, where $|\mathbf{V}|$ is the size of the vocabulary, to represent a group of policy segments. The vector generated from supervised methods represent the categories defined by the expert annotations, while the vector generated from the unsupervised methods represents the topics based on the prior belief of how the vocabulary is distributed among the topics.

Similarities Between Categories and Topics Each category and each topic is now represented as a set of words. We evaluate the similarity between a given category and a given topic with a naïve yet practical approach. We first pick the top N most significant words for each set and calculate the size of the intersection between the two sets divided by the size of the union of the two sets. In our experiments, to ensure our ability to comprehend results during evaluation, we set N to 10.

Results and Discussion

In this section, we present the results of the top 10 words for both logistic regression and non-negative matrix factorization. We then visualize the category-topic closeness matrix with a color map and discuss the relationships between categories and topics.

Top 10 Words from Logistic Regression

In our experiment, we divided the OPP-115 corpus into training set (75) and test set (40). We evaluate our linear models with 5-fold cross validation. Logistic regression (0.75) performs slightly better compared with Support Vector Machines (0.73) in the overall micro-F1 score. In our models, we used L1 penalty which leads to sparsity since we would like to evaluate the top coefficients of the model.

We built a classifier for each category using logistic regression, which generated the best results on our test set. The top 10 words for each category extracted from logistic regression are shown in Table 1. Although we built the classifiers individually, the top vocabularies for the categories barely overlap. This indicates that different data practice categories have different vocabularies that capture their characteristics. With this observation, we look for relationships with topics extracted with our unsupervised method by comparing their vocabularies.

Top 10 Words from NMF

In our experiment, we set the number of topics, K, to 12. This is because in the ideal case, we hope that there will be a one-to-one mapping. The 12 topics in Table 2 that NMF generates appear to align with some of the topics defined in the annotation scheme. Topic1 shares vocabulary with First Party Collection, Topic8 shares vocabulary with policy change, and Topic3, 5, 7 share vocabulary with specific audiences. The policy-specific phrases are not as dominant compared to the vocabularies generated by the LDA model such as generating a topic with multiple vocabulary related to a specific policy like "honda" for honda.com. This observation indicates value in further analyzing the relationship between NMF topics and expert-defined categories.

Relationship Between NMF Topics and Expert-Defined Categories

Figure 1 shows a heat map of the intersections between vocabulary sets for categories and topics, with numbers representing the ratios of the size of the intersection over the size of the union. Recall that both sets involved in this computation consist of the top 10 highest-weighted words. Figure 1 also shows the result as we increase the number of topics. From the figure, four types of alignments can be observed that suggest different interpretations of the relationship between the respective categories and topics: first, a unique one to one mapping, which is the ideal case; second, multiple categories aligned to the same topic; third, a single category aligned to multiple topics; and forth, no alignment between categories and topics.

One-to-One Mapping An one-to-one mapping of category-to-topic would be ideal for aligning categories and topics. However, there is no such alignment in Figure 1. We can easily imagine how we can assign categories from topics if each topic is aligned to one category.

One-to-Many Mapping An example for this kind of mapping is the alignment between International and Specific Audience and Topic 3, Topic 5 and Topic 7 as in Figure 1 when K equals 12. The name of the category already suggests that it captures at least two audience concepts. These two different audiences match the vocabulary of the three topics, child,

Categories	Vocabularies
First Party Collection/Use	use, collect, demographic, address, survey, service, number, customize, improve, contest
Third Party Sharing/Collection	party, share, sell, disclose, company, advertiser, behalf, provider, partner, public
User Choice/Control	opt, unsubscribe, disable, choose, choice, consent, setting, option, wish, agree
User Access, Edit and Deletion	delete , profile , correct , account , change, update, section, access, removal, request
Data Retention	retain , store , delete , deletion , database, participate promotion, send friend, record, information long, remove
Data Security	secure , security , seal , safeguard , protect, ensure, compromise, encrypt, advertiser set, unauthorized
Policy Change	change , change privacy , policy time , current , policy agreement, update privacy, post, decide, update august, notice
Do Not Track	signal , track , track request , respond , browser, advertising for, disable, respond track, track setting, platform visting
International and Specific Audiences	child , california , resident , european , age, parent, childrens, safe harbor, 13, parental
Introductory/Generic	collectively , privacy , overview , hard , 2015, policy cover, explain, collect use, relationship, appilcation
Privacy contact information	com , 800 , question , health privacy , contact, write, street, feedback, info worldnow
Practice not covered	health , searchable , license agreement , partner privacy , database, criterion, use various, restritions, textual, request link

Table 1: Vocabulary for each category from logistic regression, the words are sorted in descending order from left to right according to their weights

european, california. This kind of alignment is also straightforward when labeling policy segments with topic models. The policy segment can be labeled as International and Specific Audience if the policy segment is similar to all Topic 3, Topic 5, and Topic 7 in terms of vocabulary distribution. It also suggests that the topic modeling would allow for automated detection of more vocabulary specific categories.

Many-to-One/-Many Mapping The many-to-one/-many mapping is more complicated to interpret. The mapping may occur due to two reasons. One, the categories co-occur frequently in policy segments so a topic may align to two categories. One example of this is evident for K=12. First Party Collection/Use and Third Party Sharing/Collection are both aligned to Topic 1. As we calculated using the segments in the corpus, the ratio of Third Party Sharing/Collection given the occurrence of First Party Collection/Use is 0.23 and the ratio of First Party Collection/Use given Third Party Sharing/Collection is 0.29. These two numbers are the highest in the co-occurrence matrix.

A second potential explanation is that topics with multiple category alignments are general topics consisting of vocabulary present in all the categories they are aligned with. For example, for K=12, Topic 2 contains terms such as *privacy policy* and *policy*. User Access, Edit and Deletion, Policy Change, and Introduction/Generic are all aligned to Topic 2. This kind of mapping may cause confusion if we were to label privacy policies with topic modeling.

We experimented with K equals 6, 12, and 24 to see how different quantities of topics affect the relationship between categories and topics. By increasing the number of topics in topic models (K), we might be able to find a mapping. This happens when we increase K from 6 to 12, as shown in Figure 1. User Choice/Control do not show any alignment when K=6, but it is aligned to topic 11 when K=12. Therefore, adjusting the number of topics separates larger topics into smaller ones. It also allows more combinations of topics. Although the scenario is not visible in the figure, we can imagine that multiple categories align to the same set of topics but as we increase K the combinations of the categories become different.

No Mapping Although aligning categories and topics seems to be a promising direction in many cases, there exist some categories that are not aligned with any of the topics derived from topic models. This suggests two directions. First, some categories may not be identifiable solely by their vocabularies. The structure of the text, the structure of the policy segments, and how the categories correlate with each other are all possible features to further improve automated labeling of the privacy policy segments. Second, the performance of current supervised approach is not good enough. For example, there are only 22 instance in our training data set for Data Retention and hence it is never aligned to one of the topics even if we increase the number of topics.

Conclusion

Privacy policies are usually verbose and often difficult for users to read and understand. Even for experts, creating annotations for privacy polices requires significant effort. If we can create reliable unsupervised labeling approaches, we would be able to automatically analyze and label the enormous amount of privacy policies online without requiring expert annotations. Privacy policy documents share similar structures and vocabularies due to their purpose, which should make them amenable to automated analysis. In this paper, we presented experiments leveraging unlabeled data for the labeling of paragraph segments in privacy policies. Our results show the existence of meaningful mappings between topic models and categories defined by experts. We

Topics	Vocabularies
topic1	personal information, personal, information, share, collect, information collect, party, use, use information, service
topic2	privacy policy, policy, privacy, apply, practice, policy apply, 2015, question, link, policy update
topic3	child, 13, age, information child, age 13, child age, knowingly, knowingly collect, parent, child 13
topic4	identifiable, personally identifiable, personally, identifiable information, non personally, information, non, share personally, share, sell
topic5	harbor, safe harbor, safe, harbor program, european union, european, union, department commerce, certification visit, certification
topic6	browser, cooky, computer, ip, ip address, device, web, use, track, visit
topic7	california, privacy right, market purpose, california privacy, market, right, resident, direct market, california resident, direct
topic8	change, post, change privacy, update, time, policy time, notice, time time, post change, page
topic9	law, legal, require, legal process, right, require law, comply,process,disclose, safety
topic10	email, address, mail, email address, contact, provide, send, number, receive, request
topic11	advertise, ad, serve, advertisement, 'party, party advertise, visit, practice, opt, serve ad
topic12	security, protect, secure, unauthorized, unauthorized access, access, measure, physical, reasonable, maintain

Table 2: Vocabulary for each topic extracted from NMF, the words are sorted in descending order from left to right according to their weights

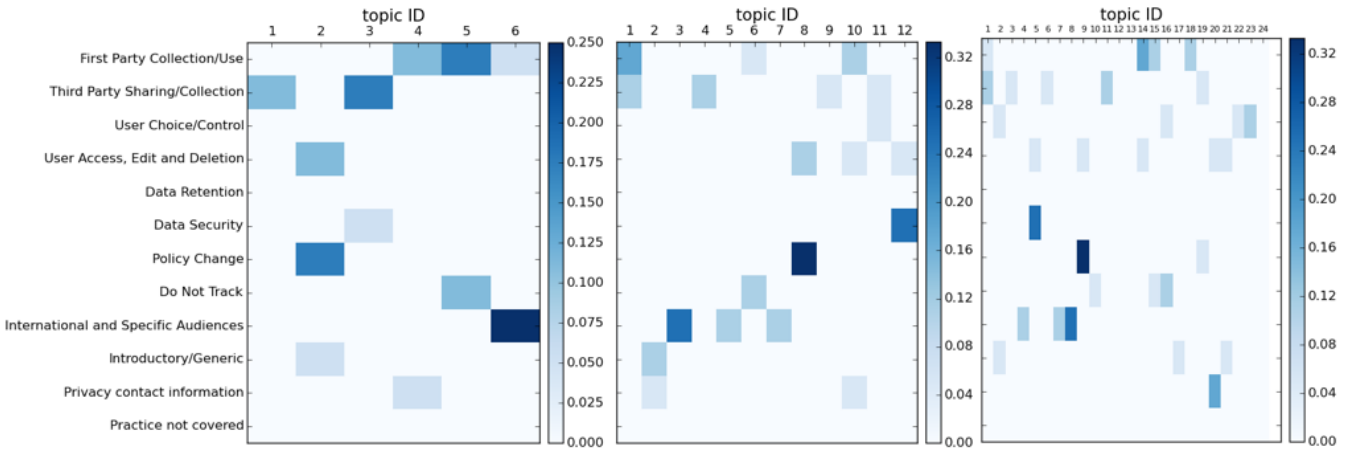

Figure 1: Color map of category/topic intersection

are in the process of fully evaluating how the proposed approach performs compared to supervised methods, but the proposed approach and our results indicate the possibility of achieving our goal by leveraging unlabeled data. In future work, we will further evaluate our approach by comparing it with existing supervised methods.

Acknowledgements

This work was funded by the National Science Foundation under grants CNS-1330596. The authors also wish to acknowledge all members of the Usable Privacy Policy Project (www.usableprivacy.org) for their contributions.

References

Ammar, W.; Wilson, S.; Sadeh, N.; and Smith, N. A. 2012. Automatic Categorization of Privacy Policies: A Pilot Study. Tech report CMU-ISR-12-114.

Breaux, T., and Schaub, F. 2014. Scaling requirements extraction to the crowd: Experiments with privacy policies. In *Int. Requirements Engineering Conference (RE'14)*. IEEE.

Chundi, P., and Subramaniam, P. M. 2014. An approach to analyze web privacy policy documents. In *KDD Workshop on Data Mining for Social Good*.

Costante, E.; den Hartog, J.; and Petković, M. 2013. What websites know about you: Privacy policy analysis using information extraction. In Pietro, R. D.; Herranz, J.; Damiani, E.; and State, R., eds., *Data Privacy Management and Autonomous Spontaneous Security*, volume 7731 of *Lecture Notes in Computer Science*, 146–159. Springer.

Ramanath, R.; Liu, F.; Sadeh, N.; and Smith, N. A. 2014. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, ACL '14, 605–610. ACL.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the*

LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50. ELRA.

Reidenberg, J. R.; Breaux, T.; Cranor, L. F.; French, B.; Grannis, A.; Graves, J. T.; Liu, F.; McDonald, A.; Norton, T. B.; and Ramanath, R. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30:39.

Sadeh, N.; Acquisti, R.; Breaux, T. D.; Cranor, L. F.; Mcdonalda, A. M.; Reidenbergb, J. R.; Smith, N. A.; Liu, F.; Russellb, N. C.; Schaub, F.; et al. 2013. The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about.

Schaub, F.; Breaux, T. D.; and Sadeh, N. 2016. Crowdsourcing privacy policy analysis: Potential, challenges and best practices. *it–Information Technology*.

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. 2016a. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics, Aug 2016*. ACL.

Wilson, S.; Schaub, F.; Ramanath, R.; Sadeh, N.; Liu, F.; Smith, N. A.; and Liu, F. 2016b. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, 133–143. International World Wide Web Conferences Steering Committee.

Wilson, S.; Schaub, F.; Ramanath, R.; Sadeh, N.; Liu, F.; Smith, N. A.; and Liu, F. 2016c. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th World Wide Web Conference*, WWW '13, 133–143.

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. USENIX Association.