Towards Automatic Classification of Privacy Policy Text

Frederick Liu Shomir Wilson Peter Story
Sebastian Zimmeck Norman Sadeh

December 2017
CMU-ISR-17-118
CMU-LTI-17-010

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This study was supported in part by the NSF under grants CNS-1330596, CNS-1330214, and SBE-1513957. The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation. The views and conclusions contained 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.

Keywords: privacy, machine learning, classification, cnn, neural network, privacy policy

Abstract

Privacy policies notify Internet users about the privacy practices of websites, mobile apps, and other products and services. However, users rarely read them and struggle to understand their contents. Also, the entities that provide these policies are sometimes unmotivated to make them comprehensible. Recently, annotated corpora of privacy policies have been introduced to the research community. They open the door to the development of machine learning and natural language processing techniques to automate the annotation of these documents. In turn, these annotations can be passed on to interfaces (e.g., web browser plugins) that help users quickly identify and understand relevant privacy statements. We present advances in extracting privacy policy paragraphs (termed *segments* in this paper) and individual sentences that relate to expert-identified categories of policy contents, using methods in supervised learning. In particular, we show that relevant segments and sentences can be classified with average micro-F1 scores of 0.79 and 0.70 respectively, improving over prior work. We discuss how the techniques introduced in this paper have been used to automatically annotate the text of about 7,000 privacy policies. Our discussion highlights opportunities as well as limitations associated with our classification approach.

1 Introduction

Privacy policies are intended to notify Internet users about the privacy practices that are applicable to their data. Various legal regimes around the world require that website operators, app publishers, and other data processors post a notice on how they gather and share users' information [12]. This requirement results in a large number of privacy policy documents that most users, however, are unlikely to read. In fact, it was estimated that a user would need to spend at least 181 hours per year to read through all privacy policies for the services they use [5]. It is our goal in this study to make the contents of privacy policies more transparent by leveraging machine learning and natural language processing techniques.

Recent work to annotate large numbers of privacy policies has enabled the use of automated methods toward policy analysis. The OPP-115 Corpus of privacy policies [10], along with its annotation scheme produced by legal experts, provides a springboard for such efforts. Enabled by the OPP-115 Corpus, this paper presents work on using convolutional neural networks (CNNs), logistic regression (LR), and support vector machines (SVMs) to classify policy text into one or more privacy practice *categories*, which represent topics that frequently occur in policy text. We classify policy text at two levels of granularity: sentences and *segments*, which (as defined by the OPP-115 Corpus) roughly correspond to paragraphs. We show CNNs to be competitive on this task with the other two methods, often with higher precision but lower recall. The best results show micro-F1 scores of 0.79 and 0.70 for segment and sentence classification, respectively, suggesting the practicality of tools built upon them. Internet users can benefit in terms of reading less if they are only interested in certain categories. For example, users interested in "First Party Collection/Use" only have to read 16% of the policy at the sentence level and 34% at the segment level as shown in Table 1.

2 Related Work

Prior computational work on privacy policy text used information extraction techniques to gather instances of data types mentioned in policies [3], opt-out choices [9], or answers to categorical privacy questions [1, 13, 14]. Closer to our work, one previous study approached the annotation of privacy policy segments as an alignment problem by using Hidden Markov Models [6]. Other

Category	sentence	segment
First Party Collection/Use	373 (16%)	796 (34%)
Third Party Sharing/Collection	280 (12%)	470 (20%)
User, Choice/Control	85 (4%)	164 (7%)
User, Access, Edit &Deletion	34 (2%)	69 (3%)
Data Retention	16 (1%)	15 (1%)
Data Security	54 (2%)	117 (5%)
Policy Change	23 (1%)	67(3%)
Do Not Track	7 (.3%)	12 (1%)
International & Specific Audiences	124 (5%)	194 (8%)

Table 1: Mean absolute counts and percentages of annotated tokens, i.e., words, at the sentence and segment levels per category per policy.

approaches leveraged Latent Dirichlet allocation [2] to facilitate privacy policy comprehension. Our work differs from these prior studies by our framing of the privacy policy analysis task as a multilabel classification problem. This approach is particularly appropriate because a segment of text in a privacy policy can contain information about multiple topics, such as first party collection of data and data security.

The manual annotation of privacy policies has been recognized as a serious bottleneck to modeling their contents, and various prior efforts were aimed at automating the annotation process [7, 11]. They derived motivation from the fact that human annotation is time-consuming, as multiple annotators must carefully interpret legal texts to produce reliable annotations. Some have proposed the automation of assigning category labels to policy segments [10]. Here, we are exploring the OPP-115 Corpus' use for classifying privacy practices in both sentences and segments.

For our task we make use of the Usable Privacy Policy Project's [8] OPP-115 Corpus, which contains detailed annotations for the data practices described in a set of 115 website privacy policies [10]. At a high level, annotations fall into one of ten data practice *categories*, which were developed by a team of legal experts:

1. *First Party Collection/Use*: How and why a service provider collects user information

2. *Third Party Sharing/Collection*: How user information may be shared with or collected by third parties

3. *User Choice/Control*: Choices and control options available to users

4. *User Access, Edit, & Deletion*: If and how users can access, edit, or delete their information

5. *Data Retention*: How long user information is stored

6. *Data Security*: How user information is protected

7. *Policy Change*: If and how users will be informed about changes to the privacy policy

8. *Do Not Track*: If and how Do Not Track signals[1] for online tracking and advertising are honored

9. *International & Specific Audiences*: Practices that pertain only to a specific group of users (e.g., children, residents of the European Union, or Californians)

10. *Other*: Additional privacy-related information not covered by the above categories[2]

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

[2] Because of its indistinct nature, we omit this category from further discussion.

Category	Vocabularies
First Party Collection/Use	use, collect, demographic, address, survey, service
Third Party Sharing/Collection	party, share, sell, disclose, company, advertiser
User Choice/Control	opt, unsubscribe, disable, choose, choice, consent
User Access, Edit and Deletion	delete, profile, correct, account, change, update
Data Retention	retain, store, delete, deletion, database, participate
Data Security	secure, security, seal, safeguard, protect, ensure
Policy Change	change, change privacy, policy time, current, policy agreement
Do Not Track	signal, track, track request, respond, browser, advertising for
International & Specific Audiences	child, California, resident, European, age, parent

Table 2: Vocabulary for each category obtained via logistic regression. Words and collocations are sorted in descending order from left to right according to their weights.

3 OPP-115 Corpus and Annotation Scheme

Privacy policies were divided into *segments*, which were roughly equivalent to paragraphs, and annotators identified spans of text associated with data practices inside of each segment. Each privacy policy was read by three annotators, who required a mean time of 72 minutes per document. In aggregate, they produced a total of 23,194 data practices.

We proceed with the observation that the text associated with each category has a distinct vocabulary, even though many of the categories represent closely related themes. Preliminarily, we applied logistic regression to identify particularly relevant words for the different categories. Table 2 shows the results. The top six words or collocations for each category show its distinctiveness.

4 Privacy Policy Text Classification

In this section we describe our procedure for labeling privacy policy text at the sentence and segment levels. Different granularity gives different results on the number of tokens annotated, which would result in different reading time if the classification results were used in downstream tasks such as simply highlighting the selected category.

4.1 Transforming OPP-115 Annotations into Labels

Annotations for data practices inside a segment can be effectively "elevated" to cover the entire segment, i.e., a segment receives a binary label for the presence or absence of each data practice category. Wilson et al. ([10]) calculated inter-annotator Kappa for segment-level labels to be 0.76 for the first two categories listed above, which comprised 61% of all data practices in the OPP-115 Corpus, and found a variety of lower and higher Kappa values for the remaining categories. For our present work, we use segment-level labels produced by a simple majority vote: if two annotators agree that a segment contains at least one data practice in a given category, then we apply that category to the segment as a label. We use a similar method to produce sentence-level labels: if

Category	Sentence			Segment			
	LR	SVM	CNN	LR	SVM	CNN	ACL16
First Party Collection/Use	.72 /.73/.73	.71/.75/.73	.77/ .71/.74	.80/**.86**/.83	.79/**.86**/.83	**.83**/.83/.83	.76/.73/.75
Third Party Sharing/Collection	.70/.72/.71	.69/.73 /.71	.74/.72/.73	.71/**.81**/.75	.70/.80/.75	**.78**/.76/.77	.67/.73/.70
User Choice/Control	.50/.60/.55	.47/.63 /.54	.69/.36/.48	.81/.58/.67	.74/**.64**/.69	**.83**/.50/.62	.65/.58/.61
User Access, Edit, & Deletion	.49/.73/.58	.49/.61/.54	.78/.41/.54	.86/**.75**/.80	.75/**.75**/.75	**.92**/.69/.79	.67/.56/.61
Data Retention	.36/.40/.38	.36/.29/.32	.45/.29/.35	.67/.29/.40	.75/**.43**/.55	**1.0**/.14/.25	.12/.12/.12
Data Security	.70/.68/.69	.69 /.66/.68	.80/.58/.67	**.82**/**.96**/.89	.79/**.96**/.87	.78/.89/.83	.66/.66/.67
Policy Change	.72/.80/.76	.74/.78/.76	.86/.59/.70	.74/.93/.82	.62/**1.0**/.77	**.86**/.80/.83	.66/.88/.75
Do Not Track	.82/.64/.72	.82/.64 /.72	.90/.64/.75	.50/.67/.57	**.60**/1.0/.75	.25/.33/.29	1.0/1.0/1.0
International & Specific Audiences	.88/.74/.81	.87/.74/.80	.89/.70/.79	**.89**/.82/.85	.85/.81/.83	.87/**.84**/.86	.70/.70/.70
Average	.69/.70/.70	.68/.71/.69	.76/.64/.69	.78/.81/**.79**	.76/.82/**.79**	.82/.76/**.79**	.66/.66/.66

Table 3: Classification results (precision/recall/F1-score) for sentences and segments using logistic regression (LR), support vector machines (SVM), and convolutional neural networks (CNN). The ACL16 results are from [10]. However, the test set is different from ours and we excluded three classes which belong to the "Other" category under the annotation schema in our average F1 score.

at least two annotators labeled any part of a sentence with a given category, we label the sentence with that category. Note that the labels are not mutually exclusive, and a segment or sentence may be labeled with zero categories or any combination of them.

4.2 Prediction Methods

For our experiment, we split the 115 policies of the OPP-115 Corpus into 80% training and 20% testing sets. Since each segment or sentence can contain information for multiple categories, we built binary classifiers for each category with three models, respectively logistic regression, support vector machines, and convolutional neural networks [4]. We used a bigram term frequency–inverse document frequency (tf–idf) pre-processor for logistic regression and support vector machines. The parameters for each model are tuned with 5-fold cross validation. The parameters for the CNN follow [4]'s CNN-non-static model, which uses pre-trained word vectors. We used 20% of the training set as a held-out development set to refine these models.

4.3 Results and Discussion

The results of segment- and sentence-level classification are shown in Table 3. Across all categories and models, we observe an average micro-F1 score of 0.79, precision 0.82 and recall 0.82, which outperforms previous results using word-embeddings as features [10].[3] Fewer tokens are annotated at the sentence level. The results at this granularity are on average about 0.1 worse than at the segment level. One explanation could be that although annotators have access to the context that surrounds a sentence (e.g., prior and subsequent sentences), our sentence-based models do not. We also observe that the CNN model favors precision while the other two models favor recall. This difference can be taken into consideration for downstream tasks with different objectives (e.g., governmental regulators might be interested in manually verifying results; hence, not miss-

[3] The data split may be different since it is not released along with the OPP-115 Corpus.

A. Third Party Sharing/Collection:

> *As Kaleida Health is a teaching facility, we may disclose your health information for training and educational purposes to faculty physicians, residents and medical, dental, nursing, pharmacy or other students in health-related professions from local colleges or universities affiliated with Kaleida Health.*

B. Data Security:

> *Chase Paymentech Solutions, LLC is committed to safeguarding the privacy and security of the information we collect.*

Figure 1: Examples of classification errors. For the first example, our models failed to detect the Third Party Sharing/Collection category. In the second example our models disagreed with our gold standard data; however, the text does appear to address Data Security.

ing instances is more important than the false positive rate). The results are consistent in F1-scores across the three models as shown in Table 3.

All three models show similar performances after careful parameter tuning, which motivates us to look at the data in more detail to find reasons for errors. For example, the OPP-115 Corpus does not contain many privacy policies of health care providers. One provider's policy is quoted in Figure 1A showing health-specific language, more of which would encourage improved performance. We also observed some errors in annotation that may have been oversights by the readers. For instance, the quote in Figure 1B was classified as a security statement, which appears to be correct, but the annotators did not recognize it as such.

During our evaluation we recognized that our classifiers' performances are also impacted by the context or lack thereof during the production of the annotations. For example, section headings were only shown to the annotators for the segment immediately following it. Features that encode context around each sentence should be investigated to avoid this problem.

Overall, our results indicate the strength of these methods toward enabling downstream tasks, such as filtering for more detailed data practices, extracting salient details to present to users, or summarization of privacy practices.

5 Conclusion

In this study we demonstrated the use of traditional and neural network models to classify text in privacy policies according to nine categories that cover important privacy practices. We believe that our results provide support for the use of segment-based annotations. At the same time we recognize that ultimately sentence-based annotations offer the prospect of finer, more detailed annotations. Further research is needed to improve the performance of sentence-based classification. Taking into account the text of adjacent sentences might help. In this paper, we point out the trade-off between number of tokens annotated and classification performance between different granularity of segmenting the policy. For future work, we are developing tools to help reduce the level of effort required from Internet users to understand privacy policies. This includes packaging the results of our analysis in the form of browser plug-ins that summarize key statements extracted from the text of privacy policies, as well as the exploration of question answering functionality to answer people's privacy questions.

References

[1] Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. Automatic categorization of privacy policies: A pilot study. Technical report, Carnegie Mellon University, 2012. CMU-ISR-12-114, CMU-LTI-12-019.

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

[3] Elisa Costante, Jerry den Hartog, and Milan Petković. What websites know about you: Privacy policy analysis using information extraction. In Roberto Di Pietro, Javier Herranz, Ernesto Damiani, and Radu State, editors, *Data Privacy Management and Autonomous Spontaneous Security*, volume 7731 of *Lecture Notes in Computer Science*, pages 146–159. Springer, 2013.

[4] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[5] 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):540–565, 2008.

[6] Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, ACL '14, pages 605–610. ACL, June 2014.

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

[8] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Noah A Smith, Fei Liu, Florian Schaub, et al. The usable privacy policy project. Technical report, Carnegie Mellon University, 2013. CMU-ISR-13-119.

[9] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. Identifying the provision of choices in privacy policy text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2764–2769, Copenhagen, Denmark, Sep 2017. ACL.

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

[11] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, pages 133–143. International World Wide Web Conferences Steering Committee, 2016.

[12] Sebastian Zimmeck. The information privacy law of web applications and cloud computing. *Santa Clara Computer & High Tech. L.J.*, 29(3):451–487, 2013.

[13] Sebastian Zimmeck and Steven M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium*, USENIX Security '14, pages 1–16. USENIX Association, August 2014.

[14] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shormir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. Automated analysis of privacy requirements for mobile apps. In *24th Network & Distributed System Security Symposium (NDSS 2017)*, NDSS 2017, San Diego, CA, February 2017. Internet Society.