Automatic Text Summarization of Website Privacy Policies Using Latent Semantic Analysis with Annotations

Nick Saulnier
Khoury College of Computer Science, MSc Student
Northeastern University
Seattle, Washington

***Abstract*— This article explains the need for more accessible representations of data collection and use practices, citing the difficulty average computer users have with understanding privacy policies, and the amount of content contained in the average policy document, as primary reasons for the general lack of awareness concerning data practices. The most significant work to date on analyzing, parsing, and categorizing privacy policies is discussed, focusing on the methods researchers have used to classify and label significant sections of privacy policies. Particular attention is given to the work that has stemmed from the collection and human-annotation of the OPP-115 website policy corpus. I then give a high-level overview of the field of automatic text summarization and its applications for the summarization of legal documents, describing my own efforts at summarizing privacy policy segments using Latent Semantic Analysis.**

I. Introduction

Most software applications collect and process data generated by computer users. The commodification of personal data has become one of the driving forces of economic growth and innovation in post-industrial societies; according to the European Commission, personalized data is projected to generate more than 1 trillion euros annually, or nearly 8% of the EU's GDP, by the end of 2020 [1]. And yet, most computer users remain ignorant of how the data they generate is collected and processed. This general lack of knowledge concerning the use of personal data is partly a result of the means through which data collection and use practices are communicated to users, namely the privacy policy. Privacy policy documents are notoriously long-winded and inaccessible to the average computer user. With an average length of 2,500 words, it would take a reader with a reading rate of 250 words per minute approximately 10 minutes to finish the average policy [2]. If individuals read the privacy policy of every website they visited in a year, they would spend over 200 hours reading, losing approximately $3,500 to the task [2]. Moreover, privacy policies are filled with technical and legal language that is often conveyed using ambiguous terminology and sentence structures that make it difficult to determine the exact nature of the collection and usage practices being described [3].

Efforts to distill essential information contained in privacy policies into more digestible representations have focused mainly on developing automatic systems for analyzing policy documents. Some of the most advanced research in this area has been conducted by researchers at The Usable Privacy Project—a joint research venture, affiliated with a handful of industry-leading research institutions, aimed at building on recent advances in the areas of natural language processing and privacy engineering to increase the accessibility of privacy policies. Using a corpus of 115 website privacy policies, curated by examining google search trends across a sub-sampling of top-level website sectors, researchers at The Usable Privacy Policy Project were able to semi-automate the process of extracting essential information describing data collection and use practices using a combination of supervised learning with annotations and deep learning with convolutional neural networks [4]. Examples of privacy policies segmented using their models are available at: https://explore.usableprivacy.org/. Building off of the work done on the OPP-115 corpus, Harkous et al. devised a deep learning system, Polisis, for segmenting previously unseen privacy policies according to a taxonomy of 10 high-level categories and 122 fine-grain classifiers. On top of their policy parser, Harkous et al. built a bot that answers free-form user questions regarding a given privacy policy [5].

As a follow-up to research proposed by Zimmeck et al., I built a prototype for an automatic text summarization system to further collate text segments from privacy policies that have been annotated with the machine learning classifiers developed by researchers at The Usable Privacy Project. To retain important details from annotated policy segments, I built a vocabulary of the 100 most important terms for each of the ten categories outlined in the annotation scheme developed at The Usable Privacy Project. I was able to build these vocabularies by scraping and analyzing text spans from the OPP-115 corpus annotations. I then used these vocabularies to extract the most topical sentences from each of the policy segments present in a given pre-segmented policy document. While the prototype works on pre-segmented policies, further work is needed to evaluate the summarized text blocks and tweak the sentence selection and combination algorithms to ensure the highest possible readability of summarized policy segments is achieved.

II. Previous Work on Automating Privacy Policy Analysis

The most significant efforts towards automating the analysis of website privacy policies have stemmed from the work done on collecting and annotating the OPP-115 corpus. As a starting point, the researchers that compiled the corpus iteratively devised, with the help of a small group of privacy experts and legal scholars, an annotation scheme to capture data collection and use practices at a level of granularity that generally describes a set of common practices that can be identified by a distinct vocabulary. The annotation scheme consists of ten categories and data collection/use practices, shown in Figure 1 with their frequencies of occurrence, mean, median, and coverage in the corpus [4]:

Category	Freq.	Mean	Median	Coverage	Fleiss' Kappa
First Party Collection/Use	8,956	78	74	.27	.76
Third Party Sharing/Collection	5,230	45	39	.21	.76
Other	3,551	31	25	.24	.49
User Choice/Control	1,791	16	13	.08	.61
Data Security	1,009	9	7	.05	.67
International and Specific Audiences	941	8	6	.07	.87
User Access, Edit and Deletion	747	6	5	.03	.74
Policy Change	550	5	4	.03	.73
Data Retention	370	3	2	.20	.55
Do Not Track	90	1	0	.01	.91

Figure 1. OPP-115 annotation categories and statistics [4]

Once the annotation scheme was devised, researchers had graduate law students annotate each policy in the corpus according to the categories in the scheme. Each policy was annotated by three different annotators. The annotations, which were consolidated and scored by the researchers after the completion of the annotation process, consist of a category tag, attribute or sub-category describing a particular data practice, and the text span where the particular data practice is outlined. Over 23,000 data practices were identified by the annotators in over 102,000 text spans.

To verify the distinctiveness of each category, researchers applied logistic regression to the text spans in order to identify the most topical words in each category. The results are shown below in Figure 2 [6]:

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

Figure 2. OPP-115 Category Vocabularies [6]

The policies were then split into subsets used to train and test logistic regression, SVM, sequence labeling, and convolutional neural network models for predicting and labeling privacy policy segments, a process detailed in two research papers [4][6]. To train their models, the researchers broke down the training dataset into 3,749 policy segments and represented them as dense vectors. Each policy segment was then assigned a binary vector with labels corresponding to the annotation scheme categories. To improve the category labeling precision of their models, only segments labeled as belonging to a specific category by two or more annotators were labeled as that category in the binary vectors.

Their initial efforts for policy segment labeling tested logistic regression, SVM and sequence labeling HMM models. They found that SVM models produced better results for nearly every category, with higher and more consistent overall precision and recall scores [4]. Logistic regression models produced recall scores comparable to the SVM models, but had inconsistent precision scores. The SVM models had micro-average precision and recall scores of 0.66 and 0.66 respectively, while the logistic regression models had scores of 0.53 and 0.65.

In their follow-up research, the researchers refined their logistic regression and SVM models by using a bigram term frequency-inverse document frequency (tf-idf) pre-processing of the input text for topic selection, tuning the models with 5-fold cross validation. The models' category labeling capabilities were trained and tested on sentence and segment-level text sizes. The switch from dense word-embedding vectors to sparse tf-idf matrix pre-processing of input text using bigrams increased the average F1 scores of the SVM and logistic regression models for segment-level labeling by 0.12 and 0.18 respectively, although it is unclear whether the use of sparse matrices over dense word-embeddings, or bigrams over single words, had a greater affect on the improved results. The CNN models were trained using pre-trained word vectors, presumably with the vocabularies discovered from each section via linear regression. The CNN model training followed the CNN-non-static approach, in which the pre-trained word vectors are updated during training to maximize labeling precision and recall capabilities. The CNN models' average F1 scores for segment-level labeling of 0.75 were slightly lower than both the SVM and linear regression models' scores. All three models produced lower F1 averages on sentence-level labeling tasks, indicating that the models produce more accurate labeling when provided greater amounts of context.

Further improving on the labeling capabilities of ML models, Harkous et al. leveraged the OPP-115 annotations to train their classifiers to "detect fine-grained labels of privacy [policy] segments" [5] that fall under the ten data practice categories defined by the OPP-115 annotation scheme. Their automated policy analysis framework, Polisis, consists of data, machine learning, and application layers. In the data layer, a policy document is scraped with headless chrome from a given URL; while the presentation of policies varies from site to site, the researchers found that most online policies, even those with dynamic content, are loaded with the page, enabling easy and efficient scraping. After a policy is scraped, a segmenter performs an "initial coarse segmentation by breaking down the policy according to the HTML and <p> tags," which is followed by a subsequent unsupervised segmentation step that uses the GraphSeg algorithm to generate fine-grained, semantically coherent segments. The machine learning layer consists of two components: an unsupervised stage for building custom word-embedding vectors from a corpus of 130k privacy policies collected from apps on the Google Play Store, and a supervised stage that utilizes transfer learning with the word-embedding vectors built in the unsupervised stage to train

neural networks in privacy text-classification. The OPP-115 annotations were used to train CNN classifiers in the supervised stage to accurately assign groups of word vectors in the dense word-embedding vectors category classifications. The CNN classifiers predict one or more high-level categories for each policy segment, refining these predictions with lower, sub-category attribute-level predictions. In total, the researchers were able to identify and assign segment-level classifications to 122 fine-grained privacy classes across the ten OPP-115 data practice categories. In an improvement on the classification models created by Zimmeck et al., the Polisis classifiers achieved an average F1 score of 0.84 against the human-annotated OPP-115 policies. The application layer takes free-form and structured user queries concerning the data practices of a given policy and outputs responses generated from information in segments tagged as relevant to the given query.

III. A Brief Overview of Automatic Text summarization and its applications for Legal Documents

Text summarization techniques fall into two general categories, defined by the methods through which summaries are created. Abstractive techniques work by first discovering the abstract topics that define a text or set of documents, a process known as topic modeling. Abstractive summarization models then take the discovered topics and use them to generate new sentences that accurately convey the central thematic content of the input corpus [7]. In contrast, extractive summarization models work by identifying the most important sentences in a document or corpus of documents, and produce a summary by concatenating these sentences together [7]. While abstractive summarization techniques have the potential to outperform extractive techniques, they typically make use of deep learning models and are more difficult to implement [7].

As such, most of the literature on automatic text summarization has been focused on extractive techniques [7]. This is particularly true of existing research on the automatic summarization of legal documents [9]. Efforts to summarize legal texts have also focused almost exclusively on single document summarizations, which is, according to Merchant and Pande, partly a result of the semantic uniqueness that characterizes individual legal documents [8]. Some of the more common extractive techniques that have been applied to legal texts include graph-based techniques built off of the LexRank scoring algorithm, latent Dirichlet allocation topic modeling algorithms, and singular value decomposition techniques based on latent semantic analysis algorithms in conjunction with tf-idf scoring [9].

While website and application privacy policies do exhibit significant amounts of diversity in terms of their content, language, levels of ambiguity, and format, they can nevertheless be categorized and classified according to a distinct set of vocabularies describing common data practices, as was shown by the researchers who compiled the OPP-115 corpus. Likewise, just as the Polisis framework improved precision and recall scores for the classification of policy segments by leveraging unsupervised learning methods on a very large corpus of privacy policies, there is great potential for the application of abstractive, deep learning summarization techniques to privacy policies, both on the single document and multi-document levels.

IV. A Starting Point for the Automatic Summarization of Privacy Policy Documents Using Latent Semantic Analysis

Because of the time constraints I was operating within, I decided to implement an automatic text summarization system for privacy policies that had already been segmented and classified by the models developed at The Usable Privacy Policy Project. While it would have been better to work with policies classified by the more accurate Polisis framework, I did not have access to a sizeable corpus of documents that have been analyzed by Polisis. Over 7,000 machine annotated and segmented policies are available for viewing at https://explore.usableprivacy.org/, in addition to the original 115 human-annotated documents. As a starting point, I wrote a scraper using Puppeteer/Headless Chrome that extracts highlighted policy segments from a given policy at https://explore.usableprivacy.org/. Each highlighted section corresponds to one of the ten categories outlined in the OPP-115 annotation scheme.

I then downloaded the original OPP-115 annotations made by graduate law students. The annotations are organized by policy into 115 separate CSV files. Each row of these files corresponds to a single annotation for a given category and one or more sub-categorical attributes, each with the span of text that the annotator applied the category/attribute classification to. Category/attribute annotations that were tagged by only a single annotator are identified as "not-selected," while annotations with a consensus of two or three annotators appear multiple times. To make the complete collection of annotations easier to parse, I concatenated all 115 CSV files into a single CSV. I then created a dictionary with an element for each of the ten annotation scheme categories, and a list of all attribute/text spans for each category in the collated CSV file. I then preprocessed each text span not labeled as "not-selected" by removing all stop words, converting all words to lower case, removing all non-inter-word punctuation, and stemming all words in each span. I then generated tf-idf scores for all unique stems in the text spans of each section using the scikit-learn feature extraction library, selecting the top-100 stems from each category. These top-100 lists formed the category vocabularies I used to score policy segments from a given pre-segmented policy. I also performed the same operation on each categories' attributes in hopes of weighting the tf-idf scores of terms that occur with more common attributes more heavily, but I was not able to fully implement this idea.

I then turned my attention to processing the segments of the input policy, performing the same cleaning operations on each pre-labeled text segment as I did for the annotation text spans. Given that I was able to compile a vocabulary of topics for each category using the annotations that were used to train the models that classified the policy segments I was working with, I decided that a simple extractive summarization approach using latent semantic analysis and tf-idf scoring might produce adequate results. As per Merchant and Pande,

latent semantic analysis (LSA) "is a completely automated unsupervised statistical-algebraic summarization technique which follows an extractive approach towards analyzing documents and finding hidden semantic relations among words and sentences of a text" [8]. While LSA algorithms can be computationally expensive when applied to large datasets, they produce decent results and are fairly easy to implement. LSA uses a matrix-reduction technique of the input matrix—usually an m x n matrix, where m is the number of terms in corpus vocabulary and n is the number of sentences in the preprocessed input—known as singular value decomposition (SVD). SVD decomposes the input matrix into three new matrices:

1) Matrix U – an m x r matrix, where m represents words and r represents concepts or topics [8].

2) Matrix Σ – an r x r, non-negative diagonal matrix of singular values that represent the strength of concepts in descending order, across the diagonal [8].

3) Matrix V – an n x r matrix, where n is the number of sentences in the document or corpus and r represents topics discovered [8].

The three matrices are calculated according to the following equations:

$$A^*A = V\Sigma^*\Sigma V^*$$

$$AV = U\Sigma$$

Figure 3. SVD Formulas [8]

I calculated the td-idf scores for all terms in the category vocabularies as they appeared in the input policy segments. I then generated a matrix V for each section's input text tf-idf scores. For each term in the matrix V for each sentence, if the tf-idf value for that term was not zero, I weighted the term's tf-idf value by summing it with its tf-idf value as calculated from the annotation text spans. This weighting accounts for the importance of specific terms as they appear in the annotations used to train the models that parsed the input policies. Thus, terms that are more correlated with the classification of a given category are given more weight when determining the most topical sentences for a policy segment labeled as belonging to that category. I then summed the weighted tf-idf scores for each term in each sentence for each category's text input and selected the three sentences with the highest sums, concatenating them together in descending order to create a segment summarization.

V. Future Work

Given the time constraints I was working within, I was unable to tweak my system's parameters to maximize ROGUE-1 scores for each summarized segment. I actually didn't even have time to calculate ROGUE-1 scores at all, which would be a necessary feature for improving the accuracy of summarizations. That being said, my goals for this project were to get a prototype up and running that could be built on and tested with future work.

Some fairly easy-to-implement changes could be immediately applied to my system to improve the quality of its summarizations. First, the tf-idf values computed for the category vocabularies as they are in the OPP-115 annotations need to be normalized between 0 and 1. This might help smooth out some of the tf-idf summation spikes and large min-max ranges I noticed when scoring sentences. The attribute tf-idf scores for each category also need to be recomputed, normalized, and applied as weightings to the tf-idf scores for the annotation text spans. This may improve the ROGUE-1 scoring of larger policy segments. As it currently exists, the summarizations for each category segment are limited to three sentences, regardless of segment size. This parameter should be calculated dynamically and further tweaked to find a suitable balance between ROGUE-1 scores and summarization length. The number of topics extracted from the annotations for each category should also be changed to reflect the number of unique topics that exist in the annotations for each category. While my algorithm concatenates the highest scoring sentences in descending order, different orders should be tested to see how this parameter affects ROGUE-1 scoring.

While LSA is an easy algorithm to work with, especially when annotation data is present, it is by no means the most accurate algorithm for extractive text summarization. Graph-based, latent Dirichlet allocation, and semantic chain algorithms should also be tested to see if they produce better results. It would also be interesting to apply abstractive summarization on policy segments summarized using extractive techniques to see if this gives more readable summaries. This would potentially require training convolutional neural networks on a very large corpus of segmented privacy policies.

VI. Conclusion

I have described the most significant literature that has been produced on the parsing, analyzing and classifying of privacy policy documents. The work done at The Usable Privacy Project, and on the Polisis framework, leveraging the OPP-115 website policy corpus has paved the way for future research and inquiry into the subject. I have also provided a high-level overview of automatic text summarization and its applications for legal documents. In my own work, I demonstrated how the OPP-115 corpus annotations can be used to produce extractive summarizations of pre-segmented privacy policies with Latent Semantic Analysis.

References

[1] Vasudha Thirani and Arvind Gupta, "The Value of Data," World Economic Forum, Sep. 22, 2017. [Online] Available: https://www.weforum.org/agenda/2017/09/the-value-of-data/. [Accessed May 20, 2020].

[2] Aleecia M. McDonal and Lorrie Faith Cranor, "The Cost of Reading Privacy Policies," I/S: A Journal of Law and Policy for the Information Society, July 2008. [Online] Available: https://lorrie.cranor.org/pubs/readingPolicyCost-authorDraft.pdf. [Accessed June 22, 2020].

[3] Reidenberg et al., "Ambiguity in Privacy Policies and the Impact of Regulation," Journal of Legal Studies, 2016. [Online] Available: http://www.cs.cornell.edu/~shmat/courses/cs5436/reidenberg.pdf. [Accessed June 22, 2020].

[4] Zimmeck et al., "The Creation and Analysis of a Website Privacy Policy Corpus," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 2016. [Online] Available: https://www.usableprivacy.org/static/files/swilson_acl_2016.pdf. [Accessed May 20, 2020].

[5] Harkous et al., "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning," Proceedings of the 27th USENIX Security Symposium, Baltimore, MD, USA, August 2018. [Online] Available: https://www.usenix.org/system/files/conference/usenixsecurity18/sec18-harkous.pdf. [Accessed June 22, 2020].

[6] Zimmeck et al., "Towards Automatic Classification of Privacy Policy Text," Carnegie Mellon University Technical Report, June 2018. [Online] Available: https://sebastianzimmeck.de/liuEtAlClassification2018.pdf. [Accessed May 20, 2020].

[7] Julian S. Griggs, "TL;DR: Automatic Summarization with Textual Annotations," Graduate Thesis, Dept. of Computer Science, Princeton Univ., Princeton, NJ, 2015. [Online] Available: http://www.cs.cmu.edu/~ab/Salon/research/Julian_Griggs.pdf. [Accessed June 12, 2020].

[8] K. Merchant and Y. Pande, "NLP Based Latent Semantic Analysis for Legal Text Summarization," International Conference on Advances in Computing, Communications and Informatics, Bangalore, India, 2018. [Online] Available: https://ieeexplore-ieee-org.ezproxy.neu.edu/document/8554831. [Accessed June 12, 2020].

[9] A. Kanapala, S. Pal, and R. Pamula, "Text Summarization from Legal Documents: A Survey," Artificial Intelligence Review, June 2017. [Online] Available: https://www.researchgate.net/profile/Sukomal_Pal2/publication/318238394_Text_summarization_from_legal_documents_a_survey/links/5d3fdafd92851cd04691f6f7/Text-summarization-from-legal-documents-a-survey.pdf. [Accessed June 12, 2020].