Automatic Classification of Web and IoT Privacy Policies

Jasmine Carson
Mount Holyoke College – Computer Science
50 College St, South Hadley, MA 01075
carso22j@mtholyoke.edu

Lisa DiSalvo
Arcadia University - Computer Science
450 S Easton Rd, Glenside, PA 19038
ldisalvo@arcadia.edu

Dr. Lydia Ray
Columbus State University
4225 University Ave, Columbus, GA 31907
ray_lydia@columbusstate.edu

***Abstract*— Privacy policies, despite the important information they provide about the collection and use of one's data, tend to be skipped over by most Internet users. In this paper, we seek to make privacy policies more accessible by automatically classifying web privacy. We use natural language processing techniques and multiple machine learning models to determine the effectiveness of each method in the classification method. We also explore the effectiveness of these methods to classify privacy policies of Internet of Things (IoT) devices.**

Keywords—privacy policies, Internet of Things, machine learning, natural language processing.

I. Introduction

A privacy policy is a statement or legal document (in privacy law) that discloses some or all of the ways a party gathers, uses, discloses, and manages a customer or client's data. Personal information can be anything that can be used to identify an individual, not limited to the person's name, address, date of birth, marital status, and contact information. However, many internet users fail to read these policies, despite the essential nature of the information they contain. According to 'The Cost of Reading Privacy Policies' by Aleecia M. McDonald and Lorrie F. Cranor, the average reader spends 8 minutes reading a 2,071 word privacy policy [5]. The average privacy policy is up 58% [9] from the 2514 words Cranor and McDonald discovered in 2008, meaning an average policy is now 3964 words long, increasing reading time in turn. Due to these time costs and the confusing wording of the policies, users tend to skip over reading them, leaving themselves uninformed on the management of their data. However, the importance of privacy policies of websites, mobile apps and Internet of Things (IoT) cannot be overemphasized in this era of ubiquitous computing, when the world is interconnected with smart devices that are continuously collecting user data [11][12][13]. For this reason, there has been research into the area of automated privacy policy reading in order to help internet users find the information relevant to them, such as the work described in "Towards Automatic Classification of Privacy Policy Text" by Liu et al [4]. This research [4] classifies privacy policies of various websites.

In this paper, we explore two different research questions:

- How do different machine learning algorithms perform in classifying the privacy policies of websites?
- If we train a machine learning classification algorithm with annotated privacy policies of websites, how will that algorithm perform to classify the privacy policies of Internet of Things (IoT) devices?

In this paper, we use various machine learning models on the OPP-115 Corpus dataset [10] to classify sections of privacy policies by topic and compare their performances. We also explore expanding these models to the privacy policies of Internet of Things (IoT) devices such as smart home devices.

II. Related Work

As mentioned previously, our work is based on "Towards Automatic Classification of Privacy Policy Text," in which Liu et al found the segment-based classification of text was a workable method for automated classification, segments being sentences or collections of sentences that appear together in the privacy policy [4]. Similarly, it was found by [8] K.M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh that machine learning could be used to identify parts of privacy policies that offer users choices, such as opting in or out of a practice, and classify these choices by topic [8]. These and our own work (described in this paper) build on the ideas outlined by The Usable Privacy Project [7]. The goal of this project [7] is to extract relevant privacy features from natural language privacy policies to assist internet users in acquiring information about their privacy. Similarly, Chundi and Subramaniam discuss the possibilities of an automatic privacy policy annotator [2]. W. Ammar, S. Wilson, N. Sadeh, and Smith perform a study on automated text classification for privacy policies [1]. While not specific to privacy policies, Yoon has studied sentence classification using convolutional neural networks (CNNs) [3].

These papers focus mainly on website privacy policies, since less research has been done on the privacy policies of Internet of Things (IoT) devices. A framework for working with the privacy policies of IoT devices has been developed by Onu, Kwakye, and Barker, describing the essential information that these policies must contain in order to be useful, such as data identifiability and data access [6].

III. Dataset

In our research we used two datasets. First, we used the OPP-115 Corpus in order to compare effectiveness of different machine learning algorithms in the automatic classification of the privacy policies of this dataset. The OPP-115 Corpus is a collection of 1,010 privacy policies from the top websites ranked on Alexa.com. The privacy policies in the dataset were retrieved in December 2013 and January 2014. Annotation of this dataset places sections of text into one of nine categories. The categories are as follows:

1. First Party Collection/Use
2. Third Party Sharing/Collection
3. User Choice/Control
4. User Access, Edit and Deletion

5. Data Retention
6. Data Security
7. Policy Change
8. Do Not Track
9. International and Specific Audiences

In order to answer our first research question, mentioned in the Introduction, we trained and tested 5 different machine learning algorithms based on this annotated dataset.

In order to answer our second research question (described in Introduction), we used a small dataset of IoT privacy policies, annotated by ourselves. The IoT annotation dataset was compiled from six IoT privacy policies: Ecobee Smart Devices, Nest Smart Devices, Rachio Smart Sprinkler, Amazon Echo, Google Home, and Fitbit, as researched and compared in [11] by Perez et al.

The data we used for our models was the selected text and category label from every annotation of the OPP-115 corpus, and those from the IoT policy dataset. Our annotation process consisted of highlighting segments from each policy and assigning it to one of the nine categories, significantly simpler than the annotation scheme of the OPP-115 Corpus. The OPP-115 Corpus dataset contains 19646 annotations, while the IoT devices dataset contains 257 annotations.

IV. Methods

To prepare the labeled texts, we performed the following steps:

- We used the Natural Language Toolkit library to remove stopwords (common words such as "the" that do not add meaning to a sentence).
- We then vectorized the texts using Scikit-Learn using the following two models:
  - a simple bag of words model
  - a term frequency-inverse document frequency (TF-IDF).

We split the OPP-115 corpus annotations 80-20 for training and testing. For the IoT privacy statements, we chose to train on the entire OPP-115 corpus and test on the IoT dataset.

The bag of words model counts the frequency of every word in the dataset for each sample, relying on the idea that samples with high frequencies of the same words are likely to be in the same category. TF-IDF also counts word frequencies, but goes a step further, changing the tallies of words based on the number of samples they occur in, since a word that appears in many samples, like a stopword, is considered to have less power to distinguish the meaning of one sample from another.

We use four machine learning approaches to classify the data: logistic regression, a Naïve Bayes classifier, support vector machines, and a neural network. The first three are implemented using Scikit-Learn, and the last with the spaCy library, a neural network for natural language processing. We split our data 80% for training and 20% for testing.

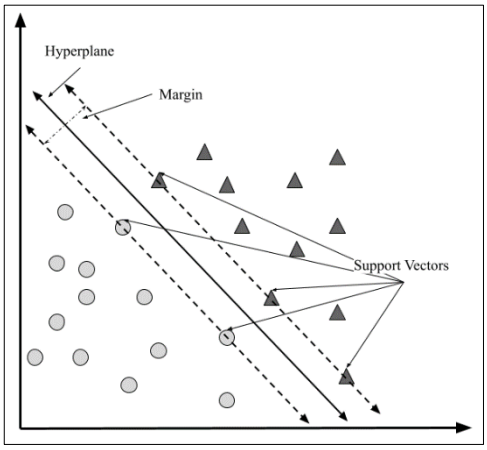

Fig. 1. Visualization of Support Vector Machine

A logistic model predicts the probability of a given piece of data belonging to a category. The equation used for logistic regression is:

$$h_\theta(x) = P(Y = 1|x) = \frac{1}{1+e^{-\theta^T x}} \qquad (1)$$

where x is the feature vector and Y is the label. The values for are optimized by gradient descent.

A Naïve Bayes classifier is based on the Bayes' Theorem on conditional probability as follows:

$$P(A \mid B) = P(B \mid A) * P(A)/P(B) \qquad (2)$$

In our program, this is:

$$P(\text{category} \mid \text{sample words}) = P(\text{sample words} \mid \text{category}) P(\text{category})P(\text{sample words}) \qquad (3)$$

We calculate this probability for each word in the sample in question and multiply them, then compare to the probabilities for the sample belonging to other categories.

Finally, a Support Vector Machine (SVM) finds an equation for dividing the data into categories and optimizes the dividing line by maximizing the distance from the line to the nearest data points, as shown in Fig. 1. This is done to make as many of the feature weights as possible zero, so that only the most important features have a weight.

With the spaCy neural network, each segment of text has every category labeled as positive or negative. Upon finalizing the count of categories, the program then outputs how many categories were identified in the data parsing. Then, using a feedback loop, the program begins to iterate recognizing words based on the training/testing data. We used an existing medium-sized English language model. This model by default has POS tagger, Dependency parser and Named entity recognition functionalities. We used the native spaCy bag of words model as well as the convolutional neural network (CNN) that the library provides.

After obtaining the statistics on the accuracy of these methods, we chose to make a vocabulary comparison between the two datasets, comparing the most common words of each category in each, to investigate further.

V. Results

We found that our models were all over 70% accurate on the dataset created from the OPP-115 Corpus, using the bag of words method, and both logistic regression and support vector

machines performed better with TF-IDF, but the Naïve Bayes classifier performed less well, as can be seen in Table I. The spaCy neural network achieved 84% accuracy, outstripping the other methods in this task.

We believe that the lower accuracy of the CNN is due to its mean pooling: the pooling of layers loses a lot of valuable information and it ignores the relation between the part and the whole. So, the positive-negative categorical matching in the training and testing method of the data set information can be lost when running the CNN function, hence producing a lower precision score.

We believe that the lower accuracy of the CNN is due to its mean pooling: the pooling of layers loses a lot of valuable information and it ignores the relation between the part and the whole. So, the positive-negative categorical matching in the training and testing method of the data set information can be lost when running the CNN function, hence producing a lower precision score.

Shown in Table II, the models performed less well on the IoT privacy policies, achieving a maximum of 67% accuracy using logistic regression with TF-IDF. The neural network was only half as accurate as it had been on the first dataset.

We found two possible reasons for these stark differences in accuracy. First, our annotation scheme was not the same as that used to produce the OPP-115 Corpus, which may have made the data from the IoT privacy statements different in form from the rest. Second, we suspected that the two sets of privacy policies had differences in vocabulary, since they pertain to different topics.

In order to check how the difference in the vocabularies of the two datasets. as mentioned before, we created lists of the most common words in each category for both datasets and compared them manually (to avoid finding a mismatch between, for example, two tenses of a verb).

Table III displays the overlapping terms in the top 10 words of a category (10 from OPP-115, 10 from the IoT policies) as a percentage, where 0% means there are no words in common, and 100% means that all words are in common.

TABLE I. ACCURACY OF MODELS ON OPP-115 CORPUS DATA

	Bag of Words	**TF-IDF**
Logistic Regression	0.79	0.81
Naïve Bayes	0.78	0.73
Support Vector Machine	0.77	0.81
spaCy Bag of Words	0.84	(N/A)
spaCy CNN	0.73	(N/A)

TABLE II. ACCURACY OF MODELS ON IOT PRIVACY POLICY DATA

	Bag of Words	**TF-IDF**
Logistic Regression	0.63	0.67
Naïve Bayes	0.59	0.58
Support Vector Machine	0.62	0.64
spaCy Bag of Words	0.42	(N/A)
spaCy CNN	0.43	(N/A)

TABLE III. VOCABULARY COMPARISON BY CATEGORY

Category	**Overlap in Top Ten Words**
First Party Collection/Use	70%
Third Party Sharing/Collection	75%
User Choice/Control	20%
User Access, Edit and Deletion	40%
Data Retention	20%
Data Security	90%
Policy Change	70%
Do Not Track	0%
International and Specific Audiences	50%

In analyzing the lists of words, we found that some of the vocabulary differences came from word choice, such as the words "change" and "update" in the category of User Access, Edit, and Deletion. Others, such as the 0% overlap in the Do Not Track category, stemmed from the differences in functionality between websites (the source of the OPP-115 Corpus) and IoT devices, which cannot receive or handle Do Not Track requests, a website-only practice. Furthermore, due to the small number of privacy policies in the IoT dataset, some of the most common words proved to be the product names, which of course could not overlap with the website data. Similarly, the IoT data did not contain the word "cookies," which was common in the website data, but did contain words such as "device" and "voice," since some devices listen for verbal commands, and these words were not found in the website data.

VI. Conclusion

We conclude, given the high accuracy of our models, that automatic classification of the text of privacy policies is a feasible method for making privacy policies easier to read, so that internet users can identify only the text that is relevant to them. We also see that, if this classification is to be extended to the privacy policies of things other than websites, the same

framework will not be sufficient, due to its website-focused construction, and more data on IoT device privacy policies is needed.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation REU Program under Grant No. 1950416.

REFERENCES

[1] W. Ammar, S. Wilson, N. Sadeh, and Smith, Automatic categorization of privacy policies: A pilot study. Carnegie Mellon University, CMU-ISR-12-114. School of Computer Science, Carnegie Mellon University, Pittsburgh, 2012.

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

[3] K. Yoon. Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882, 2014

[4] F. Liu, F. Wilson, P. Story, S. Zimmeck, and N. Sadeh, Towards Automatic Classification of Privacy Policy Text. Carnegie Mellon University, CMU‐ISR‐17‐118R, Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2018

[5] A.M. McDonald, L. F. Cranor, The Cost of Reading Privacy Policies. I/S: A Journal of Law and Policy for the Information Society, 4 (3), 543-568, 2008

[6] E. Onu, M. K. Kwakye, and K. Barker, Contextual Privacy Policy Modeling in IoT., Intl Conf on Cyber Science and Technology Congress, Calgary, AB, Canada, 2020, 94-102.

[7] N. Sadeh, A. Acquisti, T.D. Breaux, L.F. Cranor, A.M. McDonald, J.R. Reidenberg, N.A. Smith, F. Liu, N.C., Russell, F. Schaub, S. Wilson, The usable privacy policy project. Carnegie Mellon University, CMU-ISR-13-119, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2013

[8] K.M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. 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.

[9] Schwab, Reading privacy policies of the 20 most-used mobile apps takes 6h40. IntoTheMinds, 2018

[10] S. Wilson, F. Schaub, A.A. Dara, F. Liu, S. Cherivirala, P.G. Leon, M. S. Andersen, S. Zimmeck, K.M. Sathyendra, N.C. Russel, T.B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, The Creation and Analysis of a Website Privacy Policy Corpus. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1, Berlin, Germany, Association for Computational Linguistics, 1330-1340, August 2016

[11] A. J. Perez, Z. Sherali, J. Cochran. A review and an empirical analysis of privacy policy and notices for consumer Internet of things. In Security and Privacy, Volume 1, Issue 3, March 2018, retrieved from https://doi.org/10.1002/spy2.15 on June, 2021

[12] A. Perez, S. Zeadally, N. Jabeur, Investigating security for ubiquitous sensor networks. Proc Comput Sci. 2017;109:737-744.

[13] S. Zeadally, M. Badra, Privacy in a Digital, Networked World: Technologies, Implications and Solutions. Switzerland: Springer; 2015.