Automated Approach to Improve IoT Privacy Policies

Technical Report, 2018

Parvaneh Shayegh, Vijayanta Jain, Amin Rabinia, and Sepideh Ghanavati

University of Maine, Orono, ME, USA

Abstract. The massive growth of the Internet of Things (IoT) as a network of interconnected entities [18], brings up new challenges in terms of privacy and security requirements to the traditional software engineering domain [4]. To protect the individualsâĂŹ privacy, the FTCâĂŹs Fair Information Practice Principles (FIPPs) [6] proposes to companies to give notice to the consumer about their data practices, provide them with choices and give them means to have control over their own data.. Using privacy policy is the most common way for this type of notices. However, privacy policies are not generally effective due to two main reasons: first, privacy policies are long and full of legal jargon which are not understandable by a normal user; second, it is not guaranteed that an IoT device behave as it is explained in its privacy policy. In this technical report, we propose and discuss our methodologies to analyze privacy policies. By the help of this analysis, we reduce the length of a privacy policy and make it organized based on privacy practices to improve understanding level for the user. We also come up with a method to find the inconsistencies between IoT devices and their privacy policies.

Keywords: Privacy Policies · Internet of Things (IoT) · Supervised Machine Learning · Topic Modeling.

1 Introduction

Data privacy has become a major challenge for companies and social network providers, particularly in the digital world where goods and services are provided to the users on the Internet. To provide the users with reliable and efficient products (whether goods or services), companies inevitably make use of the users personal data. For example, the music industry (e.g. Soundcloud, YouTube, etc.) uses the search history of its users to provide them with the type of music they are more interested in. In such circumstances, it is of great importance to notify the users about their data privacy risks and help them make rational decisions accordingly. In doing so, companies should present their data practices in a clear and transparent way. Data practices are referred to the individual's personal data that are collected, used or shared by companies. A practical way to notify the users about data practices is to provide them with the company's privacy policy prior to the selection of any goods or services. Such privacy policies allow

users to choose which parts of their personal data can be shared, and which third parties can have access to their personal data.

Despite the potential benefits of privacy policy statements prior to selection of goods and services, these privacy policies have some issues which should be addressed. First, such statements are often ineffective, long and complicated. Schaub el. [19] discuss the reasons behind such ineffectiveness and show that privacy policies are mostly ignored by the users. Even if privacy policies are not ignored, users cannot understand them due to the complexity of privacy policy statements. To address these shortcomings, various approaches have been suggested in the literature such as *platform for privacy preferences project* [15], *text categorization* [22] [5] [3], *unsupervised classification* [14] and *recognized ambiguity* [17]. These approaches leverage privacy analysis techniques to help users understand privacy policies more clearly. Privacy analysis techniques also provide a solid background for further process of privacy policies e.g. making them concise and comprehensive. Second, if we assume that a privacy policy is short enough to be read and it is completely understandable, the problem related to privacy policies still remains unchanged, especially for IoT devices. For IoT devices, two important challenges exist: first, it is not guaranteed that privacy policies reflect exactly practices on personal or sensitive data. These inconsistencies between privacy policies and what the code of the device appears to do is one of the challenges with respect to privacy policies. Some of the previous research address similar problem between mobile applications and their privacy policies [26] [20] but the inconsistency between an IoT device and its privacy policy is still remained as a research gap.

The other challenge is that when IoT devices transfer personal information among each other, what does it happen to the transferred data? For example, if device A sends personal information to device B, and then the user removes personal information from device A, what happens to the personal information which is stored on device B? Thus, not only we should check the consistency between privacy policy of device A with its code, but we need to check the code of device B as well. That is, the analysis of privacy policies and mapping them to the devices' code is essential when we want to improve privacy policies.

In this technical report, we introduce a method to analyze privacy policies and to reduce the length of a privacy policy which is the first step in identifying inconsistencies. We, also, highlight different parts of a privacy policy based on its topic and its data practices. The applicability of this solution is tested using empirical data from privacy policies of 147 IoT devices and mobile applications. In addition, we propose a process-based approach to address the problem of IoT devices inconsistencies with privacy requirements. Testing the applicability of this approach is not discussed in this technical report.

The rest of the paper is organized as follows: First, in Section 2, we give an overview of previous work on privacy policy challenges and analysis processes. In Sections 3 and 4, we discuss the detailed approach to extract privacy policies and to analyze privacy polices based on classification algorithms. We, then, compare three classification methods with each other to find a solution for privacy policy

analysis with higher precision in Section 5. After analyzing privacy policies, in Section 6, we remove any unnecessary sentence which is not related to the user and make the privacy policy shorter and concise. Next, we introduce a method to categorize privacy policy parts based on their topic and also we present a visualized model for privacy policies in Section 7. Finally, in Section 8, we discuss the future work and conclusion of the paper.

2   Related Works

A conventional and primary approach to deal with lengthy privacy policies was the Platform for Privacy Preferences Project (P3P) [15]. This approach aimed to help users make rational decision about their privacy on websites without reading them. The P3P approach automatically retrieved privacy practices from websites that had standard format [15]. This approach was used for a long time with the Internet Explorer (IE) browser. Although the P3P approach had many advantages (e.g. helping users understand privacy policies in an organized and simple way), it also had major problems: it limited users access to websites, it did not recognize all privacy practices and it only worked with websites which defined their privacy policies using the P3P format [16].

Wilson et al. [22] argue that automated annotation of privacy policies is required prior to addressing any problem related to privacy policies. They produce a dataset consisting of 115 privacy policies (267K words) with manual annotations for 23K fine-grained data practices. The dataset includes privacy policies of top five search queries in Google web browser. They use crowdsourcing technique to annotate privacy policies and to investigate whether the data is collected or shared by a third-party. They apply logistic regression, Support Vector Machine (SVM) and Hidden Markov Model (HMM) to predict the categories (i.e. labels) of privacy policy segments and compare the performance, precision, recall and F1-scores of these three classifiers. The results shows that the SVM approach has a superior performance among the others; The HMM and the logistic regression perform similarly. Although this study provides a thorough comparison of three techniques in annotating privacy policies, it does not consider ambiguities which exist in privacy policies.

Reidenberg et al.[17] describe that website privacy policies often contain ambiguous language that undermines the purpose and value of privacy notices for website users. They develop a theory of vague and ambiguous terms which could address privacy policies' ambiguity. They propose a method to score parts of privacy policies based on their ambiguity. The method classifies an ambiguity in "share", "collect", "retain" and "use". They apply their proposed method to privacy policies of companies providing different types of services. The comparisons indicates that food industry companies, e.g. Costco, have privacy policies that are more ambiguous than companies that are subject to some form of regulation, e.g. Bank of America. Their approach, however, only detects ambiguities but it does not take a further step in solving these ambiguities within privacy

policies. The authors, however, have stated that a combination of their method with privacy policy annotation may solve this problem.

Constane et al. [5] suggest another approach to solve the problem of reading long privacy policy statements. They use text categorization and machine learning to categorize paragraphs of privacy terms to assess their completeness with a grade. They claim that this method helps users inspect a privacy policy in a structured way and only read those paragraphs that interest them. Following their study, Zimmeck et al. introduced "privee" [25] - a method to analyze privacy policies which integrates ConstaneâĂŹs classification method with WilsonâĂŹs crowdsourcing method mentioned above. In doing so, the privacy policy is returned to the user if the privacy analysis results are available in the crowdsourcing repository. Otherwise, the privacy policy is automatically classified and then is returned to the user.

Ammar et al. [3] come up with a text categorization method to prepare a pre-processed privacy policy for further Natural Language Processing (NLP) and Machine Learning (ML) analysis of privacy policies. They use crowdsourced annotations of around 60 privacy policies as a dataset to train the classifier. They employ logistic regression - a classic high-performance probabilistic model- to label privacy policy documents. In doing so, they classify privacy policies into two classes: absence or presence of a concept. They experiment different concepts to determine whether the company has access to personal data and users can cancel, terminate, or delete their accounts. They mention that their proposed method can be improved by increasing the number of crowdsourced privacy policies and experimenting more concepts.

Ramanath et al. [14] introduce unsupervised alignment of privacy policies using Hidden Markov Models. They apply their method to a dataset consisting of crawling privacy policies of around 1000 unique websites. To label each part of the privacy policy and title them accordingly, they use Amazon crowdsourcer and train their dataset based on the crowdsourcers output. In the next step, they apply HMM to this dataset where each HMM corresponds to a topic within the privacy policy. This HMM method uses a group of keywords to distinguish the topic, such as *Art*, *Business* or *Game*. The evaluation shows that the approach is more effective than previous classification techniques such as clustering and topic modeling. However, the shortcoming of this approach is that privacy policies are labelled manually which makes it unfeasible for larger datasets.

Shayegh and Ghanavati [12] proposed a method to analyze privacy policy by annotating them and highlighting each term based on its role in text. They used these annotated privacy statements, to generate a graph-based view of privacy policies and show data practices in a better way to users. This graph is also useful for adding legal requirements to privacy policies. In addition, a novel way to extract notices and choices from this graph is introduced in their paper.

Zimmeck et al. [26] develop a method to check the compliance of Android applications and privacy requirements. To extract privacy requirements of each Android app, they analyze privacy policy of that app using a machine learning method. Their research has two major parts: (1) privacy policy analysis and (2)

mobile application analysis. In order to prepare the dataset for this study, they crawled Google play store and downloaded around 18000 free Android apps. To analyze their privacy policies, they use a group of classification methods: to train the data, they manually annotated 120 privacy policies as a corpus; then, they picked a box of keywords as a feature for classification. They prototype various classifiers and conclude that support vector machine and logistic regression have the best performance. As for the mobile app analysis, they design an app analysis system based on Androguard -an open source static analysis tool. Using Androguard, they could investigate the calls from other services to check which part of information is shared with third-parties. At the end, they map privacy policy of each application with the output of this system to find the inconsistencies.

Slavin et al. [20] propose a method that links privacy policy phrases with Application Programming Interface (API) methods that produce sensitive information. Their proposed method aims at checking the compliance between the implementation of an Android application with its privacy policy. In doing so, they crawl play store and download 50 privacy policies. Then, they extracted anthologies from these privacy policies by manually annotating them. For the purpose of mapping, they propose two violations which could be found between policies' ontology and APIs: 1) weak violations which occur when the policy describes the data practice using vague terminology, and 2) strong violations occur when the policy does not describe an appâĂŹs data collection practice. They prove that privacy policies suffer from vague phrases.

3   Analysis of Privacy Policies

In this section, we discuss the process of collection and analysis of privacy policies. In the next section, we describe the detailed process for analyzing, classifying and shortening privacy policies.

As discussed in Section 1, our end goal is to identify inconsistencies between privacy policies and the IoT devices and mobile applications. Before addressing this challenge, we need to extract relevant data practices from privacy policies and develop our dataset. For this, we extract, analyze and shorten privacy policies from both the IoT and mobile applications. After extracting data practices and features from privacy policies, we can use the result for the comparison step. Figure 1 shows the steps of our approach for shortening and categorizing the privacy policies.

3.1   Collecting Privacy Policies

We collected around 150 privacy policies from both IoT devices and mobile applications by crawling Google as well as the Playdrone [11] dataset. Then, we converted these privacy policies into simple text files.

3.2   Extracting Statements

We define two classes of statements - `sensitive` and `non-sensitive`, that are present in privacy policies. To analyze privacy policies, the first step is to break a privacy policy into its sentences and to classify its parts by deciding whether they define any data practices or not.

We call a sentence as `sensitive` if it contains information about data practices and user's choices. In contrast, we define a sentence as `non-sensitive` if it does *not* define any data practices or user's choice. For example, *"Xbox One and Kinect offer easy and approachable ways to control your games and entertainment with your voice and gestures."* [23] is a sample sentence of Kinect privacy policy which is a definition and does not impact user's choices or does not notify about any data practice. In addition, the parts of privacy policies which describe privacy rules and laws are not related to the user and the user cannot make decision about them. Thus, we will label this statement as `non-sensitive`.

To classify each sentence in all the extracted privacy policies as `sensitive` or `non-sensitive`, we use machine-learning based classifier. To find the best classifier for this task, we test three most popular classification methods on 40 of the privacy policies that we have extracted. The selected classification methods are: K Nearest Neighbour (KNN) [21], Support Vector Machine (SVM) [24] and Naive Bayes (NB) [1]. We, then, compare their precision with each other and choose the most precise one to use for the rest of the privacy policies.
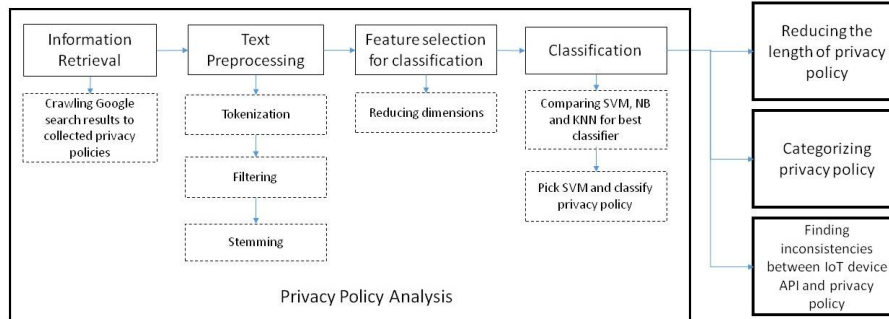

Fig. 1: Our Proposed Privacy Policy Analysis Approach

In the next section, we discuss the classification methods for the first 40 privacy policies from our dataset.

4   Classification

In this Section, we describe how we prepared the privacy policy statements to classify them as `sensitive` or `non-sensitive`. The first step before using

classification methods is to select features and train the classifier based on them. One of the common approaches to prepare a text for classification is bag-of-words technique [7]. In this technique, grammar and order of words in a text or sentence are not important. The only important factor in this model is the presence of a word and its frequency. This technique is useful in our work because we aim to use frequency of occurrence of each word as a feature for classification. One of the problems with this technique is that frequency of common words like "the", "and" and "to" is too high and this high frequency can impact the classification output. These type of words are known as stop words. In order to ensure that the these stop words do not affect the result of our classification techniques, after the information retrieval step, we pre-process our data in two steps: Tokenization and Stemming. After processing we evaluate 3 classifiers - KNN, NB, and SVM; based on the performance, we find that SVM performs better than the other two classifiers and thus we use SVM to classify policy statements. The results of classification are discussed in detail in 5

subsectionTokenization After extracting the sentences from the privacy policies, we tokenize the sentences into words and remove the stop words. Stop words are commonly preposition, connector, pronouns etc. In Example 1, we remove the words *we, may, with, our, other, of, and , for, the, in , this* and we keep this set of words *share, personal information, family, affiliated, companies, brands, purposes, described, privacy, statement.* We follow the same process for all of the extracted privacy policies. As a result, we have a set of words W which consists of these words $w_1...w_n$

Example 1.

"We may share Personal Information: With our family of affiliated companies and brands for the purposes described in this Privacy Statement."

$$W = w_1, w_2, w_3, ..., w_n \tag{1}$$

We notice that in each privacy policy around 120 sentences exist, and each sentence has 17 words on average. From these 17 words, around 2/3 are stop words which means that we have 6 keywords in each sentence and 720 keywords in a privacy policy. However, around 30% of these words are unique which means that we have 200 words per privacy policy. On the other hand, we found out that privacy policies have many words in common since the topics of all of them are very similar as we focus on privacy policies of IoT and mobile applications which makes the variety of words shorter. Therefore, we need to process these words to get a small bag of words.

4.1 Stemming

After tokenizing the sentences, we process the rest of the words by getting their root-word, a process known as stemming. Since the topics of privacy policies are not very wide, the words which are used in each privacy policy are not very distributed. As a result, a word or a different derivation of a word may be repeated

several times in each privacy policy. In Example 2, three different sentences from Kinect [23] are shown. In these three sentences "collect", "collected" and "collects" are used to explain data practices. However, all of them are different derivation of the word "collect".

Example 2.

1. We collect information on how your Kinect device and platform software are functioning.
2. Voice data may be collected to enable search and to control the console.
3. Kinect collects and uses body recognition data to enable you to control and play games.

In stemming, we keep the root of the word and remove other derivations. However, for the frequency, we consider the summation of frequencies of all of the derivations. We define a function $f(w)$ which is the frequency of word w.

To clarify this, in Example 2, three forms of collect are combined together and we keep only "collect". As a result $f(collect)$ is the summation of all three derivations of the word which means $f(collect) = 65$ for Kinect privacy policy.

Thus, we find the root of each word in W, and we remove other derivations of them. For the first 40 privacy policies, we concluded that around 700 unique words remain after stemming. Therefore, we have a vector of 700 words. This new W is the set of dimensions for classification.

$$W = w_1, w_2, w_3, ..., w_{700} \tag{2}$$

Then, we use function $f(w)$ and find the frequency of each word in W. Next, we store it in a vector F.

$$F = f(w_1), f(w_2)....f(w_{700}) \tag{3}$$

4.2   Feature Selection

After processing the data, we have a bag of words model which contains all words from privacy policies, we call W. This model is very useful to use as features for classification. However, the huge number of words in this model (around 700 words) make classification methods unfeasible. Thus, we apply a reducing dimension technique to make features set smaller. In order to reduce dimensions, we sort the words based on the sum of occurrence frequency between all sentences, and then we keep only the first 500 words. At the end, a dataset with 500 columns (extracted words) and around 5500 rows (number of sentences from 40 privacy policies) is built. We use them as our training dataset for training the KNN, SVM and NB classification algorithms. In our comparison, we find that SVM performs better than the other two classifiers and therefore we choose SVM to classify the rest of the privacy policies based on `sensitive` and `non-sensitive` categories.

5   Evaluation of Classifiers

In this section, we compare the performance of SVM, KNN and NB on our dataset in 5-fold cross-validation. Text classification has been used in different fields of study. One of these fields is to analyze privacy policies as we discussed in Section 3. In our approach, we aim to predict the label of each sentence in a privacy policy based on pre-defined labels which are `sensitive` data and `non-sensitive` data.

NB Classifier The Naive Bayes classifier is one of the simplest and popular classifier which is used in many applications. It uses a probabilistic model to find the distribution of different terms and uses this distribution for classification. In text classification, NB has two different models to classify the text. We use the model that takes frequency of the words into account instead of the model which uses binary values for presence or absence of the words.

SVM Classifier Support Vector Machines (SVM) is widely used for text classification problems. This method is one of the linear classifier approaches which finds the best vector to separate the items of class A from class B based on their features. In this paper, we use single SVM classifier as we only have two classes `sensitive` sentences and `non-sensitive` sentences.

KNN Classifier K-Nearest Neighbour classifier uses a distance measure to classify text. This method puts every two sentences in one class if they are close to each other [21]. We use cosine similarity to calculate distance between each two sentences [10].

Classification Step Two metrics are widely used for evaluation of classifiers: *precision* and *recall*. However, in our proposed approach, we use *True Negative Rate* as the third metric. In our approach, precision is the fraction of sensitive data which are labeled correctly among all sentences which are labeled sensitive with the classifier. Equation 4 shows the equation for calculating *precision*. *Recall* is the fraction of sensitive data which are labeled correctly among all sentences which are actually sensitive and it is shown in equation 5. True Negative Rate is the fraction of non-sensitive data which labeled correctly among all data which are actually non-sensitive and is shown in equation 6.

   In these equations, TP (True Positive) is the number of sentences which are sensitive and predicted as sensitive, FP (False Positive) is the number of sentences which are not sensitive but are predicted as sensitive, TN (True Negative) is the number of non-sensitive data which are predicted correctly and FN (False Negative) is the number of sensitive data which are predicted as non-sensitive.

$$Precision = TP/(TP + FP) \qquad (4)$$

$$Recall = TP/(TP + FN) \tag{5}$$

$$Negativerate = TN/(TN + FP) \tag{6}$$

As mentioned in Section 3, we randomly picked 40 of the privacy policies that we extracted to analyze these three classifiers and selected the one that has better *precision*, *recall* or *True Negative Rate*.

In our analysis, we found that *recall* for all of the cases is very close to 1. It means that by using our proposed approach for classification, non sensitive data is labeled as non-sensitive. The reason is that sensitive data have specific words in common. For example, collect, information, share, etc. However, non-sensitive data do not have specific keywords. Therefore, when classifier calculates the distance between a sentence and the data in sensitive class, it finds more similarity in compare with data in the non-sensitive class. For example, in sentence 1, the true label is non-sensitive since it is a definition for Alexa Interaction and it is not related to the user. However, because of the presence of words such as information, interaction and use, classifiers label it as sensitive.

According to this fact, we can conclude that we never lose any sensitive data while reducing the length of the privacy policy which we explain in Section 6.

Example 1. "Alexa Interactions" means all information related to your use of Alexa and Alexa Enabled Products.

After completing the classification step, we end up having some non-sensitive sentences with sensitive label which are measured by *precision* and *True Negative Rate*. Figure 2 presents the result for precision and true negative rate for each of the classification methods. In fact, precision shows the percentage of sensitive sentences which are labeled correctly and true negative rate shows the percentage of non-sensitive sentences which are labeled as sensitive.

Algorithms	KNN	NB	SVM
Accuracy	76%	63%	93%
Precision	88%	90%	92%
Negative-Rate	66%	68%	71%

Fig. 2: Comparison of KNN, NB and SVM based on their Precision and Negative Rate

By comparing the results, we found that SVM is a better classifier for our purpose.

Next, we used SVM to classify the rest of the privacy policies (107 privacy policies). The total sentences in these 107 privacy policies is 16822. Figure 4 shows the accuracy, precision, recall and F-Score for the rest of the extracted privacy policies and Figure **??** shows the percentage of `sensitive` and `non-sensitive` sentences.

Classifier	Accuracy	Precision	Recall	F-Score
SVM	99%	97%	99%	98%

Fig. 3: Result from Using SVM for Classifying Privacy Policies

-	Total # of Sentences	Total # of Sensitive	Total # of Non-Sensitive
Total	16822	10545 (62%)	6277 (38%)
Document	~107	~66	~41

Fig. 4: Percentage of Sensitive vs. Non-Senesitive Sentences

6 Shorten Privacy Policies

Reducing the length of a privacy policy, make them shorter and easier to be read and to be understood. In Section 3, we divided privacy policy sentences into two groups, sensitive and non-sensitive sentences. In this section, we remove the sentences with non-sensitive label by considering the fact that these sentences are not related to the user. For example, Amazon Alexa privacy policy [2] has 116 sentences in general. By removing its non sensitive sentences, the number of sentences is reduced by 38% and the number of words are decreased from 1157 to 820 words.

7 Categorize Privacy Policies

Breaking the sensitive part of a privacy policy into sub-groups makes the privacy policy categorized and well-organized. These groups help users find relevant part of the privacy policy by the data practices and the type of information which is collected, used or shared. Furthermore, users can easily understand the purpose of the company for such data practices. For example, if a user wants to know about collecting personal data, he/she can easily refer to the `collection` category and find all the relevant sentences about personal data collection.

To categorize privacy policies, we leverage Stanford Topic Modeling Toolbox [13] which is a topic modeling method to highlight each part of the privacy policy. This toolbox uses labeled Latent Dirichlet Allocation (LDA) to discover several topics in a document. To use this toolbox, we use the annotated, tokenized and stemmed text from Sections 4 and 5 as an input for the topic modeling. The topic modeling method performs better on meaningful words and thus irrelevant words must be removed. To address this problem, [13] suggests to remove those words that appear in less than four privacy policies.

We need to define the topics prior to the use of LDA topic modeling in this toolbox. The aim is to find the topics of sentences within the sensitive class based on the required data practice type. We define these sub-classes as: `Information`, `Collection`, `Sharing`, `Permission` and `Technology`.

Table 1: Definitions of Topics in Privacy Policies

Topics	Definition	Examples
Information	Any sentence which contains personal information.	Personal data, Email, Audio, Mailing Address
Collection	Any sentence which is related to collection of data.	Collect, Access, Use, Store
Sharing	Any sentence which explains sharing data with other parties.	Disclose, Share, Reveal, 3rd Party
Permission	The parts of the privacy policy that needs permission of the user to collect, share or use user's data.	Agree, Consent, Allow, Permit
Purpose	The sentences about the purpose of the data practices.	Purpose, To Provide, To Help, To Offer
Technology	Any sentence which explains technological aspects of data collection or sharing.	Cookie, Device, Session, Service

The above mentioned topics are shown in Table 1. In each row, the first word is the topic and the rest of the words are a small sample of meaningful words that we use to distinguish the corresponding topic. Table 2 shows three random sentences with their topics. As it is shown in this table, a sentence must be assigned to at least one topic but it may be assigned to multiple topics as well.

In Figure 5, we selected a paragraph from Hello Inc. privacy policy [8] and highlighted three topics based on their corresponding words to clarify how the proposed method recognizes the topics in a text. These three topics are `Information`, `Collection` and `Technology`. The distribution of these topics in the data is shown in Figure 6. The results indicate that Information is a topic which is used in more than half of privacy policies whereas Permission is a topic which is ignored by most of the privacy policies.

Also, we can use the categorized privacy policies to visualize data practices. Shayegh and Ghanavati [12] propose a method to convert an annotated privacy policy to a graph. Their proposed graph shows links between sensitive terms. We leverage their approach and combine it with our categorized text. In Figure 7, a sample sentence from Amazon Alexa privacy policy is picked. As it is shown in the graph, this sentence is related to three categories use, information and purpose.
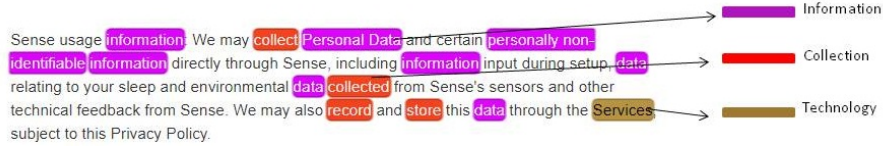

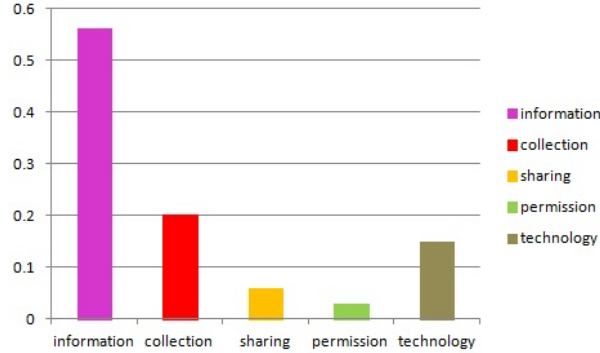Fig. 5: An Example of a Text with Each Topic

Fig. 6: The Distribution of Topics in our Dataset

8 Conclusion and Future Work

Privacy is one of the main concerns with the growth of mobile applications and IoT devices. IoT devices' users are more concern about their privacy since IoT

Table 2: Example of Three Sentences from LG Smart Home Devices' [9] Privacy Policy with Their Topics

| Sentence	Topic(s)
This information may be used to deliver Products or service which you have purchased	Information / Collect
We may collect your first and last name, mailing address, and usage data in order to track your usage of Products or Service	Collect/ Information
We generally do not share with third parties the information we receive as a result of you using the LG Smart Home Service.	share/ information
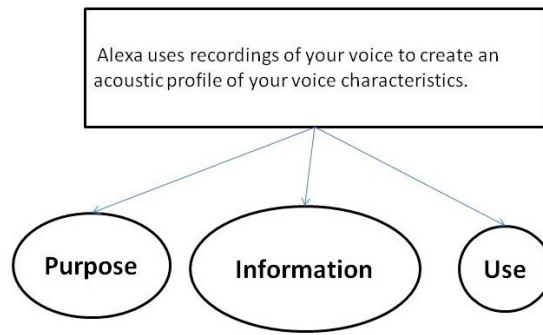
Fig. 7: Proposed Graph to Visualize Data Practices

devices can access massive amount of personal and sensitive information about the user. On the other hand, manufacturers of IoT devices are obliged to notify users about their privacy practices. The most common way for this notification is privacy policy, however, privacy policies are long, complicated and most of them do not behave exactly the same as the device does.

In this technical, we proposed a method to analyze privacy policies based on classification methods and then used the result of this analysis to make privacy policies more concise. Also, we propose a method for making privacy policies structured and categorized to help user understand privacy policies better. As a future work, we will implement this proposal.

In future, we will also develop a tool to analyze IoT device's code. By mapping the processed code to analyzed privacy policy, we aim to identify inconsistencies and propose resolution strategies for them. Identifying and resolving these inconsistencies assist developers to ensure their code to be compliant with privacy policy. In addition, having privacy policies that are consistent with the IoT device data practices, can help building trust between the users and the companies.

Our proposed approach has some drawbacks which should be addressed in future work. First, in feature selection approach, we ignore the order of the words and remove all the stop words without considering their roles in the sentence

and thus, it can result in some incorrect results. For example, in the sentence "we don't collect information", it is important to consider the word "do not" before the verb. One of the solution we aim to use in our future work is using N-gram model to store this spatial information within the text. We also keep only the most frequent words as dimensions for classification but we may lose some important none-frequent words in this way. One solution is to keep these words instead of removing them and use a classifier with higher performance.

Second, the number of IoT and mobile privacy policies that we have used for testing our approach is still not enough. In future, we aim to automate the extraction of privacy policies and build a larger dataset to ensure our results are more precise.

References

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: Text summarization techniques: A brief survey. arXiv preprint arXiv:1707.02268 (2017)
2. Amazon: Alexa terms of use (2018), https://www.amazon.com/gp/help/customer/display.html?nodeId=201809740, [Online; accessed October 2018]
3. Ammar, W., Wilson, S., Sadeh, N., Smith, N.A.: Automatic categorization of privacy policies: A pilot study (2012)
4. Anthonysamy, P., Rashid, A., Chitchyan, R.: Privacy requirements: Present & future. In: Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Society Track. pp. 13–22. ICSE-SEIS '17, IEEE Press (2017)
5. Costante, E., Sun, Y., Petković, M., den Hartog, J.: A machine learning solution to assess privacy policy completeness. In: Proc. of the 2012 ACM workshop on Privacy in the electronic society. pp. 91–96 (2012)
6. Federal Trade Commission: Privacy online: A report to congress (2016)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(Mar), 1157–1182 (2003)
8. Inc., H.: Hello inc. privacy policy (2018), https://hello.is/legal/privacy, [Online; accessed October 2018]
9. LG: Privacy policy of lg electronics usa inc, http://www.lg.com/us/privacy, [Online; accessed November 2017]
10. Liu, B., Li, X., Lee, W.S., Yu, P.S.: Text classification by labeling words. In: AAAI. vol. 4, pp. 425–430 (2004)
11. PlayDrone: Playdrone dataset (2018), https://archive.org/details/playdrone-apks&tab=collection, [Online; accessed October 2018]
12. P.Shayegh, S.: Toward an approach to privacy notices in iot. proceeding of the 4th International Workshop on Evolving Security & Privacy Requirements Engineering (ESPRE'17) (2017)
13. Ramage, D., Rosen, E.: Stanford topic modeling toolbox (2011)
14. Ramanath, R., Liu, F., Sadeh, N., Smith, N.A.: Unsupervised alignment of privacy policies using hidden markov models (2014)
15. Reagle, J., Cranor, L.F.: The platform for privacy preferences. Communications of the ACM **42**(2), 48–55 (1999)

16. Reagle Jr, J.: Designing a social protocol: Lessons learned from the platform for privacy preferences project. In: Telephony, the Internet, and the Media: Selected Papers From the 1997 Telecommunications Policy Research Conference. p. 215. Routledge (2013)
17. Reidenberg, J.R., Bhatia, J., Breaux, T.D., Norton, T.B.: Ambiguity in privacy policies and the impact of regulation. The Journal of Legal Studies **45**(S2), S163–S190 (2016)
18. Roman, R., Zhou, J., Lopez, J.: On the features and challenges of security and privacy in distributed internet of things. Computer Networks **57**(10), 2266–2279 (2013). https://doi.org/10.1016/j.comnet.2012.12.018
19. Schaub, F., Balebako, R., Durity, A.L., Cranor, L.F.: A design space for effective privacy notices. In: Eleventh Symposium On Usable Privacy and Security (SOUPS 2015). pp. 1–17. USENIX Association (2015)
20. Slavin, R., Wang, X., Hosseini, M.B., Hester, J., Krishnan, R., Bhatia, J., Breaux, T.D., Niu, J.: Toward a framework for detecting privacy policy violations in android application code. In: Proceedings of the 38th International Conference on Software Engineering. pp. 25–36. ACM (2016)
21. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in neural information processing systems. pp. 1473–1480 (2006)
22. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., et al.: The creation and analysis of a website privacy policy corpus. In: Proc. of the 54th Annual Meeting of the ACL (2016)
23. Xbox: Kinect and xbox one privacy faq (2018), https://www.xbox.com/en-US/legal/privacyandonlinesafety, [Online; accessed October 2018]
24. Zeng, Z.Q., Yu, H.B., Xu, H.R., Xie, Y.Q., Gao, J.: Fast training support vector machines using parallel sequential minimal optimization. In: Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on. vol. 1, pp. 997–1001. IEEE (2008)
25. Zimmeck, S., Bellovin, S.M.: Privee: An architecture for automatically analyzing web privacy policies. In: USENIX Security. vol. 14 (2014)
26. Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S.M., Reidenberg, J.: Automated analysis of privacy requirements for mobile apps. In: Proceedings of the Network and Distributed System Security (NDSS) Symposium. vol. 2017 (2017)