Automated Approach to Analyze IoT Privacy Policies

Alanoud Subahi[1,2][0000−0002−8642−1708] and George Theodorakopoulos[1][0000−0003−2701−7809]

[1] School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom subahiat@cardiff.ac.UK, TheodorakopoulosG@cardiff.ac.UK
[2] Faculty of Computing and Information Technology, King Abdul Aziz University, Rabigh, Saudi Arabia asubahi@kau.edu.sa

Abstract. The massive popularity of IoT devices raises new challenges for user privacy. Hence, manufacturers are obliged to notify users about their privacy practices as well as give them choices to have control over their data. Privacy policies are long and full of legal jargon, thus not understandable by average users. The problem becomes worse with IoT devices due to the ability of these devices to access sensitive information about users. Previous research has addressed problems related to websites and mobile privacy policies. However, few works focus on analyzing IoT privacy policies. In this paper, we analyze and annotate 50 IoT privacy policies to determine whether the IoT manufacturers collect personal information about the user as well as the type of such information. To ensure that we extract the correct information, we study in-depth the complicated and ambiguous sentences that average users won't understand. With our method, we aim to mimic how an ordinary person reads and understands such policies sentence by sentence. We use supervised machine learning to label the collected personal information according to its sensitivity level to either sensitive personal information or non-sensitive personal information. The high accuracy achieved by the classifier (98.8%) proves its validity and reliability.

Keywords: IoT · privacy policy · supervised machine learning · IoT privacy policy.

1 Introduction

The Internet of Things (IoT), a wide variety of smart devices connected to the Internet, is widespread both on a personal and industrial level. Smart devices range from light bulbs, switches, sensors, and kitchen appliances to TVs and wearable devices like smart watches and fitness trackers. In November 2019 Statista Research [26] projected the number of connected IoT devices to be 75.44 billion worldwide by 2025. According to McKinsey Global Institute, the financial impact of the IoT market on the global economy may reach as much as $11.1 trillion by 2025 [13].

Most of IoT devices are manufactured for personal use; therefore, they deal with a user's Personal Identifiable Information (PII) [9] all the time. Accordingly, personal user data that are collected from multiple sources have been leveraged by the manufacturers of these smart devices without a clear understanding of the associated privacy requirements. Hence, from a privacy perspective, it is essential to notify IoT users with respect to their personal data and help them make rational decisions about their privacy risks, e.g., unsolicited marketing. Although some people choose convenience over privacy as using their personal data is not a big deal for them, lots of people do have a concern about their data [21]. Many of these concerns revolve around their collected data being used by the company for targeted advertising or even sold to a third party without their knowledge.

We believe that the practical way to notify users about a company's data practices is by providing them with the company's privacy policy prior to the selection of any IoT device or services. By "data practices," we refer to the ways in which companies handle their users' personal data: collect, use, or share the data with other companies. The goal of privacy policies is for companies to describe how they handle user-collected data, and to give users a choice to select which parts of their personal data can be shared, and which third parties can have access to their personal data. Hence, IoT manufacturers are obligated to provide a sufficient Privacy Policy Agreement (PPA) and notify their users about the type of PII data they will collect while interacting with the IoT device. For example, many users wear smart watches most of the time, and thus their personal information, habits, and behavior are collected and sent to the smart watch manufacturer's cloud [24]. Consequently, it is important for the users to know what kind of personal information will be collected by the IoT device and why.

According to [12], PII can be categorized as following:

1. Sensitive PII, which refer to any information related to the user and not intended for public use or violate the user's privacy, e.g.location data.
2. Non-sensitive PII, which refer to any information related to the user and can identify his in a way that won't affect his privacy, e.g.email address.

In the ICO report [11] General Data Protection Regulation (GDPR) has clearly set the criteria for manufacturers on what data needs to be collected from the users. Additionally, companies whose business practices are found to be inconsistent with their privacy policies will face regulatory enforcement actions [8].

Although many users do care about their data practices, and they don't want any privacy breaches, most of them still ignore reading PPAs because they are too long, and some information are complex and hidden in the text [6]. However, considerable research have been done to address issues related to the websites and mobile application PPAs to help users understand these policies more clearly. In fact, researchers leverage various analysis techniques to overcome these issues, but the problem still remains particularly for IoT PPA.

Subahi and Theodorakopoulos [27] highlighted that there are some essential differences between IoT PPAs and traditional websites' PPAs due to the sensitivity of personal data transferred from the IoT device to the cloud server and vice versa. The sensitivity level of the collected data from any IoT device, e.g., the wearable device, which reveals the pattern of the life of the user, is much higher than the sensitivity level of the collected information when a user browses, searches, or even writes an email through a website. Also, Internet users need to connect to the Internet manually to search, buy, or browse the website; in contrast, IoT devices don't need any intervention from the user to initiate a connection to the Internet, except for the first time. Hence the user won't be aware when his data is collected and transferred by the IoT device to the IoT cloud.

In this research, we introduce a new method of analyzing IoT PPA texts. In particular, we are focusing on determining whether the IoT manufacturers collect PII about their end users, without asking them to read the whole PPA nor highlighting the paragraphs that refer to the data collection practices and then ask to read such paragraphs. In contrast, in our method we aim to mimic how an ordinary person reads and understands such policies sentence by sentence.

Our contribution is a tool called IoT-PPA reading, that automatically extracts from the PPA the type(s) of user's information that the IoT manufacturer collects when using their IoT devices. The main objective of this tool is to save time spent on reading long PPA text as well as reduce the effort on understanding complex and ambiguous meanings hidden in such a text. For example, if an IoT end user wants to buy a smart cam, our tool will help him to make rational decisions before using or buying any IoT device based on a prior understanding of the type of collected data, i.e. read the PPA of the IoT device and inform the user with the sensitive PII and non-sensitive PII that such a device collects.

The rest of the paper is organized as follows: Section 2 describes background and previous work in PPA analysis and their problems. In Section 3, we discuss how we collect, analyze, and annotate the IoT PPA. A brief overview of the IoT-PPA reading tool as well as a detailed description of the ten cases used to extract the features from IoT PPA in section 4. In Section 5, we develop our multi-class classifiers to classify the sentences of IoT privacy policies based on their sensitivity level. Then, we discuss the results and evaluate the performance of the tool in section 6. Finally, a summary and conclusion provided in Section 7.

2 Related Work and Background

PPAs aim to answer questions such as: what information is collected by the manufacturer? Who collects such information? How is the information collected, used, and protected? Who can access my information, and what information is being shared and with whom? A growing body of literature has examined the privacy policies of websites and mobile apps in different fields. A number of these studies have focused on evaluating the readability of PPA documents of Internet websites and mobile apps as well as assessing their language in section 2.1. In

section 2.2, we discuss various approaches that focus on annotating and categorizing the text of privacy policies. A few works have recently emerged, focusing on analyzing the IoT privacy policies of systems and devices, which we discuss in section 2.3.

2.1 Difficulties in reading privacy policies analysis

One strand of research [20], [4], [14] examines the reasons why most users ignore the PPA, what is the best time to display privacy notices to users, and why privacy policies are full of jargon and not understandable to users. While other research such as [7] suggest solutions to help users not to read the full PPA but to read only the paragraphs that belong to the categories that interest them. The previous methods aim to shorten the privacy policies so users will only read as few paragraphs as possible. However, the problems of understanding complicated, ambiguous, and hidden information [6] have not been solved.

Another strand of research has studied the readability of PPA documents within mobile environments [25], [28]. Baalous et al. [3] relied on manual testing and review to analyze the type of information collected, collection mechanisms, the purpose for collection, sharing of information, user controls and information retention period of privacy policies of cloud storage mobile applications which claim zero knowledge. However, manual testing is time-consuming despite the correct results.

In our approach, we aim to solve the previous problems in IoT privacy policies by only informing the users with the type of PII information that has been collected by the IoT manufacturer without asking them to read the full PPA or specific paragraphs. In addition, we do our analysis automatically, avoiding problems with manual analysis.

2.2 Privacy Policy annotation and text categorization analysis

Harkous et al. [10] proposed an automated framework for PPA analysis (Polisis) which automatically annotates, with high accuracy, each segment with a set of labels describing its data practices. They compare their automatic annotation with the manual annotation done by Wilson et al. [29] to prove the accuracy of their results. Although in their approach the users will read only a few paragraphs, the problem of the complexity and the difficulty in understanding the hidden meanings in such paragraphs still does not solve [6].

[2], [16]used machine learning techniques for text categorization on privacy policies to determine whether the company has access to personal data as well as if the users can cancel, terminate, or delete their accounts. Whereas Sathyendra et al. [19], [18] aimed to detect the provision of choices in the PPA as they focused on extracting opt-out instances.

In contrast, our research is different from the previous researches in the following:

1. We propose a new annotation scheme targeting IoT privacy policies in order to infer whether the IoT manufacturer collect PII information about their user or not.
2. We categorize the sensitivity level of the collected PII by the IoT manufacturer into sensitive-PII and non-sensitive PII according to the GDPR [12].
3. Our classifier works at the level of sentences instead of segments or word level as we have 31661 sentences in our 50 IoT privacy policies.
4. Our method applies ten corner cases, see section 4.3, to address and solve the problems of complicated and hidden meaning.

Reidenberg et al. [17] propose a method to score parts of privacy policies based on their ambiguity. Hence, in their study, they develop a theory of vague and ambiguous terms that could address privacy policies ambiguity. They used machine learning techniques to classify ambiguity in "share", "collect", "retain" and "use".

Our work is similar to the above study in that we also study and analyze ambiguous language but in IoT privacy policies. However, their method does not take any further steps in solving these ambiguities within privacy policies. In contrast, in our research we propose a method to solve such ambiguity, see section 4.3 for more details.

2.3 IoT Privacy Policy analysis

We find that all previous studies have focused either on; making the privacy policies of the websites and the mobile apps more readable by shortening their duration or determining whether personal information can be collected, disclosed to advertisers, or kept indefinitely. While a few works have emerged focusing on analyzing the IoT PPAs. IoT users understand that their personally identifiable information is used for some purposes. For example, smart watch users expect their data to be transferred to the company's servers to calculate their burned calories. However, they do not know the type of personal information that was transferred, nor if this information might violate their privacy.

The recent two studies related to analyzing the privacy policies of the IoT, as discussed below, does not address the problem of informing the user about the type of collected information by the IoT device's PPA. While our study is the only one who analyzes the IoT PPA and proposes a new method to inform the user about the type of PII that has been collected about him through the IoT PPA as well as categorizing such information based on its level of sensitivity to either sensitive PII or non-sensitive PII.

Shayegh and Ghanavati [22] analyzed 25 IoT privacy policies and proposed a set of new annotations. They used these new annotations to manually classify IoT PPA in order to present short notices on the IoT device's screen. As a result, they generated a graph-based view and show data practices in a better way to users. However, lots of IoT devices do not have a screen like smart switches or smart labs. While Perez et al. [15] work is different from Shayegh et al [22] in terms that they provide an analysis of the privacy practices instead of proposing

a model for the analysis.provided an analysis of privacy practices for six IoT devices and systems. They presented a review of issues related to privacy policies about the practices that manufacturers provide related to data collection, data ownership, data modification, data security, external data sharing, policy change and policies for specific audiences.

In contrast, our work is different from the previous researches in three main things: first, among studies that analyze IoT privacy policies e.g. [22], [15] the largest dataset contains only 25 privacy policies, while in our research we analyze twice as many (50 policies). Second, we propose a new set of annotations for; 1) specifying the type of information collected, i.e. if the IoT devices PPA collects user log in information, 2) categorizing the collected data either as sensitive PII or non-sensitive PII according to the general data protection regulation to the GDPR [12]. Finally, we point out that the classifier works at the level of sentences, and we have 31661 sentences in our 50 policies.

3 Collecting IoT Privacy Policies

To perform our analysis and apply our annotation scheme, we need to collect a range of IoT PPAs. We select our policies based on the popularity of the IoT manufacturers. In total, we come up with 50 different IoT PPAs, covering smart home appliances, smart kitchen appliances, smart security devices, smart wearable devices, and smart health and tness devices. See appendixA.

3.1 Annotation scheme

To annotate each PPA, we apply two phases. The first phase explains the manual annotation scheme. While the second phase explains the automated annotation scheme as following.

Manual Annotation In this phase, we manually annotate ten out of fifty IoT PPAs. We create four main annotation labels, which are **"Collect", "Sensitive", "Non-sensitive", and "Not-include"**, see Figure 1.

In addition, we create extra sub-annotations for the last three main annotations, see Figure 2. These sub-annotations help us to be more accurate regarding the type of the collected data by the IoT PPA, as per the following explanation:

1. Collect: we label any phrase or word that means "collect user information by first party" as "Collect". Notice that we only care about the first party collection, which represents the IoT manufacturer.
2. Sensitive: we label any phrase or word that means "user sensitive PII information" such as user location, user log in details, or user password information as "Sensitive". Under this annotation, we define three sub-annotations # Location,# Log in,# Password. For example, the sentence "we collect user location" is labeled as Collect, Sensitive PII-Location.

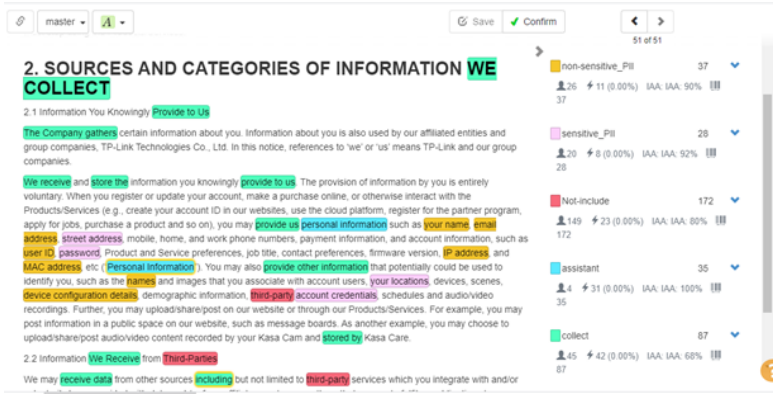

Fig. 1. An example of annotating the Tp-link PPA

3. Non-sensitive: we label any phrase or word that means "user non-sensitive PII information" such as user email address, username, or device information as "non-sensitive". Under this annotation, we define three sub-annotations # Email, # Username, # Device. For example, the sentence "you provide us with your first name" is labeled as Collect, Non-sensitive PII-username.
4. Not-include: under this annotation, we define nine sub-annotations # Negative-words, # Wrong-words,# Share-words,# Third-party,# Cookie-words,# Wrong-credentials, # Wrong-location, # Wrong-email, and # Wrong-name. We will explain how we label the text according to this annotation in section 4.3.

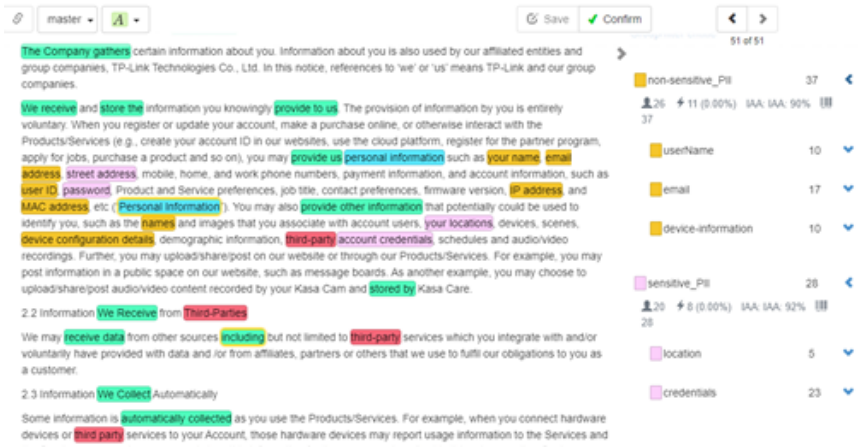

Fig. 2. An example applying the sub-annotation

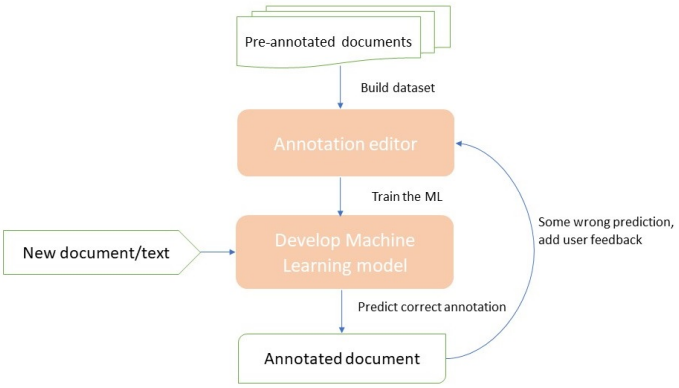

Fig. 3. The process of how to use tagtog custom ML to automate the annotation scheme

Based on the previous annotation scheme, we are ready to label the rest of the IoT PPA automatically as we explain in the next phase.

Automatic Annotation It is time consuming if we continue to annotate the rest of the 40 IoT PPAs manually; hence, we need to automate the annotation process. To do that, we use a web-based annotating tool called tagtog [1]. According to Cejuela et al. [5] illustrated how tagtog-assisted annotation can benefit manual and automatic annotation and shows a successful annotation with high accuracy.

To better use this tool, we need first to annotate manually a few documents (phase one). Second, based on the annotation scheme, tagtog will generate a model to annotate the new documents by creating a custom ML model automatically. Figure 3 shows the automated annotation process in tagtog tool. It is important to emphasis that we manually verified the annotations that tagtog produced.

After annotating all the IoT PPA documents, we extract only the labeled phrases and remove the unlabeled one. As a result, we get 31,661 labeled phrases. For example, the phrase "providing us location" is labeled as "CollectLocation-sensitive", while the phrase "you may supply us your e-mail" is labeled as "CollectEmail-nonSensitive", and so forth for the rest of the phrases. We use this dataset for training and testing our classifier as we explain in section 5. Moreover, we create five different assistant datasets for our feature extraction rules,i.e. the ten corner cases, as follows:

Dataset#1 includes phrases or keywords that represent negative meaning (neg-K), e.g. "not collect", "we don't collect", "we won't collect".

Dataset#2 includes phrases or keywords that mention a "collect" keyword without implying that any user data is being collected, i.e. wrong collect

(wc-K), e.g. "When you access your location", "to provide you with latest update".

Dataset#3 includes phrases or keywords that mention data sharing (share-K), e.g., "when you choose to share your location", "we share your personal information".

Dataset#4 includes phrases or keywords that mention third-party involvement (thirdParty-K), e.g. "we collect your third-party account information".

Dataset#5 includes phrases or keywords that mention cookies collection (cookie-K), e.g. "our cookies store your log in details".

4 Methodology

In this section, we first give a brief overview of the IoT PPA tool in section 4.1. Then we explain in detail how we create and apply ten different cases to help us extract the correct features in section 4.2. Finally, we explain how such cases can adversely affect the validity of extracting the results in section 4.3.

4.1 Overview of the IoT-PPA reading tool

Initially, the tool asks the user to provide the URL of the IoT PPA as an input. After that, the tool processes the document in order to prepare it for features extraction. The results are saved in a CSV file for later prediction. Finally, the classifier classifies the sentences of the IoT PPA into one or more of six classes according to whether it collects sensitive PII or non-sensitive PII information, as follows:

1. "CollectLocation-sensitive",
2. "CollectPassword-sensitive",
3. "CollectLogin-sensitive",
4. "CollectEmail-nonSensitive",
5. "CollectUsername-nonSensitive",
6. "CollectDevice-nonSebsitive".

Figure 4 gives an overview of the proposed method. We make out tool publicly available at(https://github.com/AlanoudSubahi/IoTPPA_Reading_Tool).

4.2 Data processing

To prepare the collected IoT PPAs for the analysis conducted in this research, we need first to pre-process the collected data. The methodology that we use includes the following steps:

1. We use "Urllib.request" module for fetching the URLs of the IoT PPA; the result of this module is a text contained HTML and XML tags.
2. To extract the HTML text only and remove all unwanted tags, we use "Beautiful-Soup2 library".

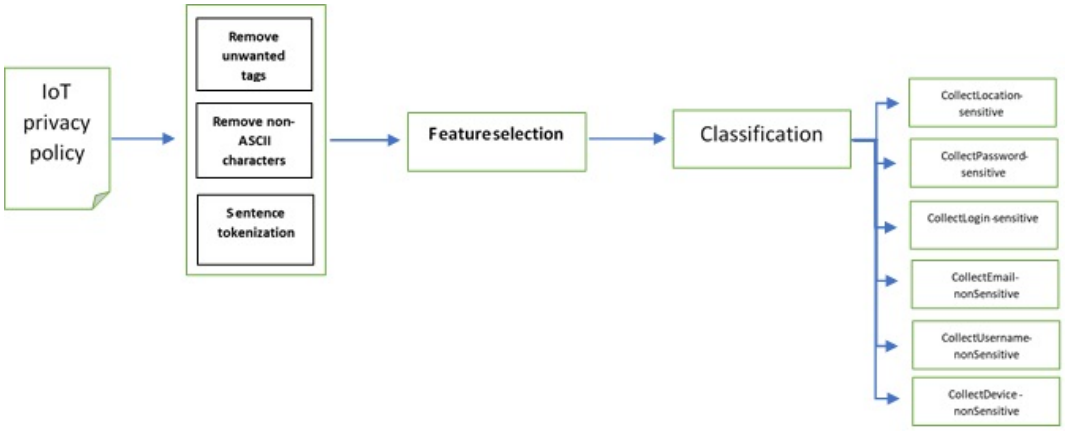

Fig. 4. Overview of the proposed method of analyzing the IoT privacy policy documents

3. We use "Regular Expressions" to remove non-ASCII characters such as punctuation and special characters.
4. The final text has been tokenized into sentences using "Natural Language Toolkit", and lower case them.

In contrast with other approaches, i.e. [23], we do not remove English stop words such as "you", "we", "they" etc. because, in our analysis, we consider the role of the party who performs the action. In total, we process 31,661 sentences from 50 IoT privacy polices.
Once the IoT PPAs are ready, we start applying our annotations scheme to each sentence. After that, we use these sentences as instances to extract the relevant features for the classification algorithm.

4.3 Extracting relevant Features

This section aims to extract from each sentence individually, whether it has one or more of the following features. We assign 1 if the feature/s exist; otherwise, we assign 0. Accordingly, we build six different functions, each of which is responsible for extracting one feature. The features are:

1. Location feature
2. Log in feature
3. Password feature
4. Email feature
5. Username feature
6. Device feature

According to the GDPR [9], PII categorized to either sensitive PII or non-sensitive PII. Therefore, the first three features are considered sensitive PII.

While the last three features are considered non-sensitive PII. In our approach, we aim to imitate how a person could understand the meaning of a sentence, i.e. knows whether the sentence collects sensitive PII or non-sensitive PII.

Before we explain our method, we must first clarify that the previous approach to finding out whether a PPA document collects personal information or not is **keyword matching:**. This method checks whether the text contains any word from the collection keywords list such as "collect", "provide", ...etc. Also, it checks whether the text contains any word from the PII keywords such as "location", "password", "username", ...etc. Hence, if the **keyword matching** method finds both keywords in the text, then the PPA collects PII about the users. Otherwise, it does not collect any PII about the user. To prove whether such a method is reliable or not, we will test it using three different examples as follows:

Example 1 if we have the sentence, "We collect your personal information such as your geographic location, email address and your device software information." The **keyword match** method will conclude that the sentence collects your location, email address, and device information because it matches the keywords. This is a positive result.

Example 2 if we have the sentence, "We collect your personal information to improve our services", the **keyword match** method will conclude that the sentence dose not collect PII about the user because it only match the "collect" keyword, and there is no word matches the PII keywords. This is a positive result.

Example 3 if we have the sentence, "We will not collect your geographic location", the **keyword match** method will conclude that the sentence collects geographic location. This is false results because the sentence does not collect any PII about the user. The reason behind this false result is that **keyword matching** method does not consider the impact of the negation words within the sentence.

Consequently, the main objective of our method is to overcome the previous false results and any similar ones due to the ambiguity of the meaning. Thus, we study in-depth all the possible cases that might affect understanding the correct meaning of such a sentence. As a result, we come up with ten different cases, each of which has its own set of rules. These rules depend on two main conditions:

1. The role of the party (i.e if its the manufacturer as a first party or the end user as a second party).
2. The position of the keywords in the sentence (i.e the collect keyword, the sensitive keyword, the negative keyword...etc).

To guarantee that we collect the correct feature(s), We should apply these cases onto each sentence in order. In Figure 5 we applied ,in order, the ten cases with its rules to illustrate the process of extracting one feature, i.e. the location feature from a sentence. We apply the same method for the rest of the features.

In the first case, we explains how we deal with the negative keyword if it is exist in the sentence. While, from the second case until the sixth case, we

explain how we address the problem of long, ambiguous, and complicated sentences. Finally, from the seventh case until the end, we explain how we treat four different type of ambiguous sentences, which imply hidden meaning of collecting information. We will now discuss each case separately:

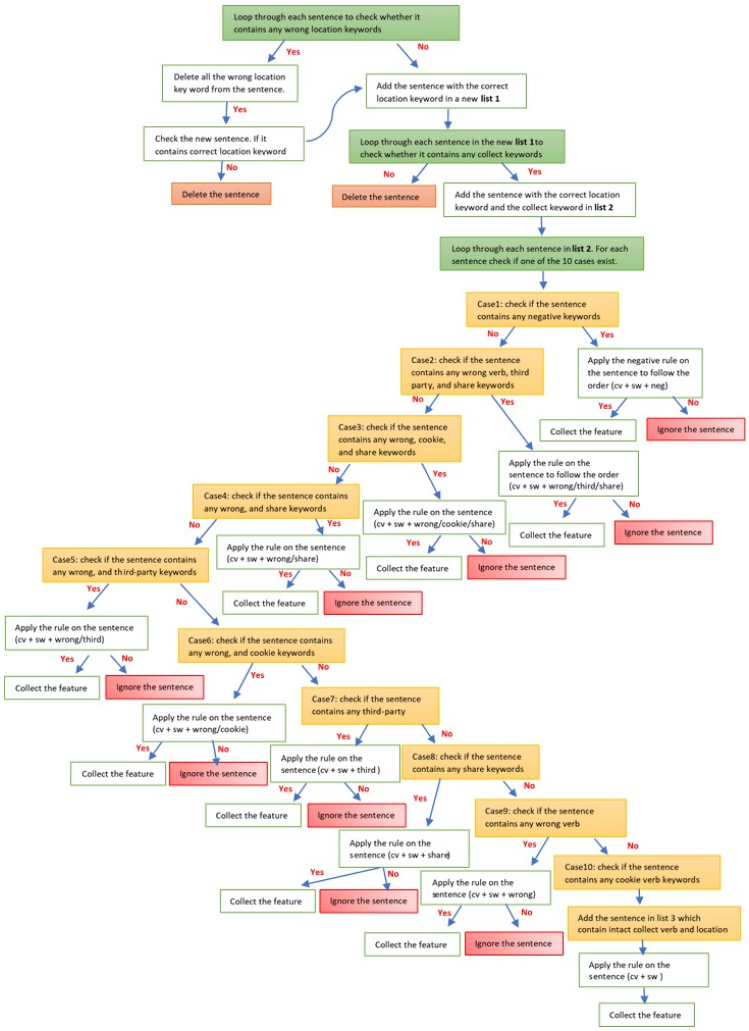

Fig. 5. An Example of how we apply the ten corner cases to extract location feature

Case 1: Negative sentences In this case we ensure that the sentence does collect users information, if so, we continue until we extract all the feature(s). Otherwise, we delete the sentence from the list. To do so, the tool loops through

the negative Dataset#1, in section 3.1 and checks whether the sentence contains any negative words (neg-K). If so, we have to identify the position of such keywords in the sentence by applying three different rules. These rules are:

1. If the position of the neg-K comes before the position of the sensitive keyword (s-K) and the collect keyword (c-K), then we ignore the sentence. For example, in the sentence, "If you do not wish to have your location recorded while taking a photo, you can turn this off at any time within the camera settings of the device". The negative phrase "you do not" comes first, then the sensitive phrase "your location", then the collect keyword "recorded". Hence, if the rule is (neg-K + s-K + c-K) or (neg-K + c-K + s-K), then we ignore the sentence.
2. If the position of the neg-K comes in between the s-K or the c-K, then we also ignore the sentence (c-K + neg-K + s-K) or (s-K + neg-K + c-K). For example, in the sentence, "We may ask you not to turn on your location". The negative phrase "not to" comes between the collect keyword "we may ask" and sensitive keyword "your location".
3. If the position of the neg-K comes after the s-K and the c-K, then we are sure that we extract the correct feature. For example, in this sentence, "This location data is collected anonymously in a form that does not personally identify you", the s-k "location" comes first, then the c-k "is collected", then the neg-k "does not",i.e., (s-K + c-K + neg-K) or (c-K+ s-K + neg-K).

Case 2: Long and complicated sentences (combination of wrong collect keywords, third-party keywords, and share keywords) In this case, we study the first type of complicated sentences, which include a combination of, wrong collect keyword (wc-K), third-party keyword (thirdParty-K), and share keyword (share-K). For example, we have this long and complicated sentence after processing the PPA of Ring manufacturer for smart doorbell[3] *"The types of personal information we obtain include: Contact information, such as name, phone number, and email; Account information, such as online password and other log-in details used to access Ring products and services; Payment information, such as name, card number,...etc"*

Initially, the average reader can be confused in understanding the type of information that the sentence collects and who is responsible for collecting it. In fact, a sentence like this is too long and complicated so the user cannot immediately understand it. However, by careful reading we can infer the following information:

1. The manufacturer of Ring obtains personal information such as password and log in details, which consider sensitive PII, as well as information such as name, phone, and email, which consider non-sensitive from the user.
2. On behalf of Ring, a third-party payment processor collects payment information from the user, such as username, card number and expiration date, and security code.

[3] https://en-uk.ring.com/pages/privacy-notice

3. Only if the user chooses to log in to her Ring account through third-party social services such as Facebook Ring will obtain her personal information such as log in details.
4. If the user chooses to share her video information via social media such as Facebook, Ring will obtain this video information from the user.

The user is only concerned about the type of personal information the IoT manufacturer collects about him, i.e. the first point only. Hence, we build our tool to handle these long and complicated sentences in order to help users understand the meaning of such complicated sentences. First, the tool checks if any word from the wrong collect keywords and any word from the third-party keywords and any word from the share keywords exists in the sentence (dataset #2, #3, #4 in section 3.1). If we find all the words, we create a list that contains the index of each word within the sentence. After that, we divide the sentence into partitions based on these indices. For the example of the sentence above, the keywords that we find are **"to access"**, **"third-party"**, and **"you share"**. Hence, the new sub sentences of the previous sentence are the following:

1. "The types of personal information we obtain include: Contact information, such as name, phone number, and email; Account information, such as online password and other log-in details used to access."
2. "Ring products and services; Payment information, such as name, card number, expiration date and security code, which is collected and stored by our third-party."
3. "payment processor on our behalf; Information we obtain from third-party."
4. "social media services (e.g., Facebook) or payment services (e.g., PayPal) if you choose to link to, create or log into your Ring account through these services (including when you share."
5. "Ring videos or content via your social media account); Information we obtain from third-party."
6. "business partners if you choose to use our Ring+ Service, such as your account ID, account name, and email address."

To guarantee that our tool extracts the correct features, we apply the following rules on each partition.

- The first rule is related to the wrong collect keyword. If any of the sub-sentences include either this rule (c-K + s-K + wc-K) or this rule (s-K + c-K + wc-K), then we collect the feature. Otherwise we ignore the sentence.
- The second rule is related to the third-party keyword. If any of the sub-sentences include either this rule (c-K + s-K + thirdParty-K) or this rule (s-K +c-K + thirdParty-K), then we collect the feature. Otherwise we ignore the sentence.
- The third rule is related to the share keywords. If any of the sub-sentence include either this rule (c-K + s-K + share-K) or (s-K + c-K + share-K), then we collect the feature. Otherwise we ignore the sentence.

By applying these three rules, we come up with the same results we previously inferred from the sentence, i.e. the first point. The results : "we obtain name", "we obtain email", "we obtain password", and "we obtain log in".

Case 3: Cookies instead of third parties Case 3 is similar to Case 2. The only difference is that we search for a cookie keyword instead of a third-party keyword. For example, we have this long and complicated sentence after processing the PPA of Google home manufacturer[4] *"Examples of how we use your information to deliver our services include: We use the IP address assigned to your device to send you the data you requested, such as loading a YouTube video;...etc."*

By careful reading, we infer from the sentence that Google home manufacturer doesn't collect any personal information. Hence, the purpose of our tool is to give us the same result. Therefore, we apply the same rules related to the wrong collect keyword and the share keywords as before. Moreover, we apply further rules related to the cookie keywords, which are either (c-K + s-K + cookie-K) or (s-K + c-K + cookie-K). As a result, we conclude that the previous sentence does not collect any personal information, which is similar to what we infer manually.

Case 4: Long and complicated sentences (a combination of wrong collect keywords, and share keywords In this case, we study the third type of complicated sentence, which only includes a combination of wrong collect keyword and share keyword. For example, we have this sentence after processing the PPA of Ezviz manufacturer[5] *"When you save and share content through EZVIZ Services, like video clips, live video streams, images, captions, and comments (Your Content), for other individuals to access,...etc."*

By careful reading, we infer from the sentence that EZVIZ manufacturer doesn't collect any personal information. We address this case just like **Case 2 and Case 3**. We divide the sentence into partitions based on the index of the wc-K and share-K. By applying the same rules related to the wc-K and share-K, we conclude that the sentence does not collect any personal information from the user.

Cases 5: Long and complicated sentences (a combination of wrong collect keywords, and third-party keywords) Case 5 is similar to Case 4, except the sentences include only a combination of wc-K and thirdParty-K.

In contrast, **Case 6: Long and complicated sentences (a combination of wrong collect keywords, and cookie keywords)**, also similar to Case 4. However, the sentences include a combination of wc-K and cookie-K. By applying our rules, we conclude that the results we obtain from the tool are similar to what we infer from previous sentences.

[4] (https://policies.google.com/privacy
[5] (https://www.ezvizlife.com/uk/legal/privacy-policy

Cases 7, 8, 9, and 10 single keyword These cases are about ambiguous sentences which contain at least one keyword. As mentioned earlier, we have already built a dataset of all possible phrases that include third-party keywords, share keyword, wrong collect keywords, and cookie keywords, during the analysis stage in section 3.1). We now explain each case separately:

Case 7 In this case,the tool checks whether the meaning of the sentence implies collecting personal information by third-party. Hence, we apply three different rules as follows to ensure that we extract the correct results.

1. If the position of the third-party-K comes between s-K and the c-K, then we collect the feature i.e. (s-K + thirdParty-K + c-K) or (c-K + thirdParty-K + s-K). For example, "we collect and use information obtained from Facebook, Google, Amazon, and other accounts you use to log in to the Services ("third-party Accounts"), such as your name, ...etc."
2. If the position of the thirdParty-K comes after the s-K and the c-K, then we collect the feature i.e. (c-K + s-K + thirdParty-K) or (s-K + c-K + thirdParty-K). For example, "we collect your email, or log in for a third-party account (like Facebook) ".
3. If the position of the thirdParty-K comes first then the c-K then the s-K, we ignore the sentence i.e. (thirdParty-K + c-K + s-K) or (thirdParty-K + s-K + c-K). For example, "When you purchase LIFX Products through the LIFX Website, our third-party provider will collect, your first and last name, email address ".

Case 8 In this case, the tool checks whether the meaning of the sentence implies collecting personal information for share purposes. In this case, we apply two different rules:

1. If the position of the share-K comes in (c-K + s-K + share-K) or (s-K + c-K + share-K), then we collect the feature. For example, "we will collect information about your exact location when you choose to share that with us."
2. if the position of the share-K is (share-K + s-K + c-K) or (share-K + c-K + s-K), then we ignore the sentence. For example, "The share information also includes the information related to you shared by other users who use the services of Mobvoi including collect location data and log information".

Case 9 In this case, the tool checks whether the meaning of the sentence implies collecting personal information when it actually didn't collect any personal information. Hence, we apply three different rules:

1. If the position of the wc-K is (c-K + s-K + wc-K) or (s-K + c-K + wc-K), then we collect the feature. For example, "We collect information that your Device sends out or receives to tailor the Services to our users in different regions, such as: geo-location, IP addresses".
2. If the position of the wc-K is (c-K + wc-K + s-K) or (s-K + wc-K + c-K), then we ignore the sentence. For example, "Include fulfilling orders

for products or services, delivering packages, sending postal mail and e-mail, processing payments, transmitting content, and providing customer service."

3. If the position of the wc-K is (wc-K + c-K + s-K) or (wc-K + s-K + c-K), then we ignore the sentence. For example, "You can access your information, including your name, or address."

Case 10 In this case, the tool checks whether the meaning of the sentence implies collecting personal information by cookie. Hence, we apply three different rules:

1. If the position of the cookie-K is (c-K + s-K + cookie-K) or (s-K + c-K + cookie-K), then we collect the feature. For example, "Other information collected automatically through the foregoing means may include your IP address, location details, cookie information, and other indicators of how you are interacting with the Services."
2. If the position of the cookie-K is (c-K + cookie-K + s-K) or (s-K + cookie-K + c-K), then we ignore the sentence. For example, "We treat information collected by cookies and other technologies as non-personal information, except where Internet Protocol (IP) addresses."
3. If the position of the cookie-K is (cookie-k + c-K + s-K) or (cookie-K + s-K + c-K), then we ignore the sentence. For example, "We use cookies,for a shopping basket or for the OSRAM login and which your browser stores."

After applying all the ten corner cases, in order, onto each sentence, we are sure that our tool extracts the correct features.

5 Machine Learning-Based Classification

To solve our classification problem, we compare several popular classification algorithms from different literature. Accordingly, we train five machine learning models i.e. Decision Tree, Linear Support Vector Machines, Random Forest, Multinomial Naive Bayes, and Multi-Layer Perceptron to classify IoT PPA texts based on (a) whether it collects sensitive PII or non-sensitive PII, (b) the type of such PII. To do this, we use the dataset that we have already created during the analysis stage (section 3.1). We randomly split the dataset into 60% for training, 20% for validation, and 20% for testing and evaluating the performance of our tool, see section (section 6). We train each of these classification algorithms using the training dataset, and we evaluate them with the following four metrics:

- **True positive (TP)** - the number of sentences that are sensitive and are correctly predicted as sensitive.
- **False positive (FP)** - the number of sentences that are non-sensitive but are falsely predicted as sensitive.
- **True negative (TN)** - the number of sentences that are non-sensitive and are correctly predicted as non-sensitive.

– **False negative (FP)** - the number of sentences that are sensitive but are falsely predicted as non-sensitive.

As is standard in the literature, from these four metrics we calculate three more: precision, recall, and F-measure. Precision (P) is the fraction of the sentences that are correctly labeled as sensitive among all sentences that are labeled sensitive by the classifier *[Precision = TP/ (TP + FP)]*. Recall (R) is the fraction of the sentences that are correctly labeled as sensitive among all sentences *[Recall = TP/ (TP + FN)]*. F-measure (F) calculates precision and recall; it takes both false positives and false negatives into consideration to evaluate the overall classification performance *[F1Score = 2 *(Recall * Precision)/ (Recall + Precision)]*. Accuracy calculates the fraction of the sentences that are predicted correctly to the total number of sentences *[Accuracy = (TP + TN)/ (TP + FP + FN + TN)]*.

Based on the results of the previous measurements, shown in Table 1, we find that all the classifiers achieve high accuracy. However, to select the best classifier, we compare the time efficiency to accomplish the task of each classifier. Hence, Multinomial Naive Bayes classifier achieves the best performance resulting in 97.4%, 97.4%, and 97.5% respectively. Also, the Multinomial Naive Bayes classifier achieves the shortest time in performing the task with 0.16 seconds for 18997 sentences. To evaluate the classifier and to ensure that it avoids

	Common Measures			
Classifier	P	R	F	time (in second)
Decision Tree	98.1%	98.1%	98.1%	0.70
Multi-Layer Perceptron	98.9%	98.9%	98.9%	5.5
Support Vector Machine	98.2%	98%	98%	68.8
Random Forest	98.4%	98.4%	98.4%	1.07
MultinomialNB	97.5%	97.4%	97.4%	0.16

Table 1. The results of all selected classifiers based on the most common measurement; precision, recall, and F1-score

over-fitting problems, we perform the following experiments:

Confusion matrix experiments To better understand the performance of the selected classifier, we create confusion matrices of the classifier in Table 2. The predicted label of the individual sentence appears in the columns while the actual label appears in the rows. For example, the actual number of the sentences that collect password information (the fifth row) is 542. However, the classifier correctly predicts 498 sentences as collectPassword-sensitive; in contrast, it predicts incorrectly that 44 sentences are collectLogin-sensitive. The overall results confirm that our classifier achieves high accuracy, and we can rely on such a classifier to classify the IoT PPA.

Compare the accuracy of the training dataset with the accuracy of the

	Predicted labels					
	cDevice-nS	cEmail-nS	cLocation-s	cLogin-s	cPassword-s	cUsername-nS
cDevice-nS	1533	0	0	0	0	0
cEmail-nS	0	453	0	0	0	0
cLocation-s	0	0	2019	0	1	0
cLogin-s	0	0	0	572	117	0
cPassword-s	0	0	0	44	498	0
cUsername-nS	0	0	0	1	0	1094

Table 2. Confusion matrix of the Multinomial classifier. Rows show the actual class of repetition and columns show the classifier's prediction. Row and column titles have been abbreviated using "c" for "collect," "s" for "sensitive," and "nS" for "nonSensitive."

validation dataset One of the methods that we use to ensure whether we have an over-fitting issue or not is comparing the accuracy of the validating dataset with the accuracy of the training dataset. As we can see in Table 3, both results are very similar; hence we conclude that there is no over-fitting.

	Multinomial classifier
Train accuracy	97.57%
Validation accuracy	97.42%

Table 3. The accuracy of the training data and the validating data

10-fold cross validation The best way to determine optimal values of hyperparameters is through GridSearchCV over possible parameter values using k-fold cross-validation on different random subsets of our labeled dataset. We use $k = 10$ where a random $(k\text{-}1)/k$ fraction of the dataset is used to train the classifier, and the remaining $1/k$ are tested for accuracy. Based on the results we set our hyperparameters as follows: alpha = 1.0, fit-prior = True, and class-prior = None.

The results of the previous experiments prove that our classifier doesn't fall in over-fitting problems.

6 Results and discussions

To evaluate the performance of our tool, we apply the trained classifier to the 20% of test dataset (i.e. 6,332 unseen sentences). The results show that the classifier classifies the sentences with high accuracy equal to 98.8%. As a result, we prove the validity of such a tool to infer whether the IoT PPA collects sensitive or non-sensitive information about the user.

7 Conclusion

In this research, we describe our approach to analyze and extract personal information from 31661 of 50 IoT privacy policies. In contrast with previous research, we don't highlight the paragraphs that refer to data collection practices, because that leaves it to the user to try understanding the hidden and ambiguous meaning of such paragraphs. Rather, our method gives the user exactly the type of information that is collected about him.

Our tool reads the IoT PPA text sentence by sentence in order to extract the correct meaning. We come up with ten corner cases; each case affects the way of understanding the correct meaning of the sentence.

Our main goal is to give the user the type of sensitive PII e.g."location" and non-sensitive PII e.g. "email address" that the IoT device collects, thus saving time and effort for the user. To fulfil our goal, we build a multi-class classifier to inform the users of the type of the collected information. Our selected classifier achieves high accuracy (98.8%) as well as high speed(e.g. 0.16 sec for classifying 18997 unseen sentences). The high accuracy results achieved by our tool prove its reliability.

Acknowledgments

The first author's work is sponsored by King Abdul Aziz University in Saudi Arabia.

References

1. The text annotation tool to train ai. https://www.tagtog.net/ (2020)
2. Ammar, W., Wilson, S., Sadeh, N., Smith, N.A.: Automatic categorization of privacy policies: A pilot study. School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019 (2012)
3. Baalous, R., Poet, R., Storer, T.: Analyzing privacy policies of zero knowledge cloud storage applications on mobile devices. In: 2018 IEEE International Conference on Cloud Engineering (IC2E). pp. 218–224. IEEE (2018)
4. Balebako, R., Schaub, F., Adjerid, I., Acquisti, A., Cranor, L.: The impact of timing on the salience of smartphone app privacy notices. In: Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices. pp. 63–74 (2015)
5. Cejuela, J.M., McQuilton, P., Ponting, L., Marygold, S.J., Stefancsik, R., Millburn, G.H., Rost, B.: tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. Database **2014** (2014)
6. Costante, E., Den Hartog, J., Petkovic, M.: On-line trust perception: What really matters. In: 2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST). pp. 52–59. IEEE (2011)
7. Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J.: The platform for privacy preferences 1.0 (p3p1.0) specification (2002)
8. Federal Trade Commission: https://www.ftc.gov/ (2020)

9. Grimes, R.A.: What is personally identifiable information (pii)? how to protect it under gdpr. https://www.csoonline.com/article/3215864/how-to-protect-personally-identifiable-information-pii-under-gdpr.html (2020)
10. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. In: 27th {USENIX} Security Symposium ({USENIX} Security 18). pp. 531–548 (2018)
11. Information Commissioner Office: https://ico.org.uk/ (2020)
12. Information Commissioner Office: What is personal data? a quick reference guide. https://ico.org.uk/media/for-organisations/documents/1549/determining_what_is_personal_data_quick_reference_guide.pdf (2020)
13. Manyika, J., Chui, M.: By 2025, internet of things applications could have $11 trillion impact. https://www.mckinsey.com/mgi/overview/in-the-news/by-2025-internet-of-things-applications-could-have-11-trillion-impact (2020)
14. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. Isjlp **4**, 543 (2008)
15. Perez, A.J., Zeadally, S., Cochran, J.: A review and an empirical analysis of privacy policy and notices for consumer internet of things. Security and Privacy **1**(3), e15 (2018)
16. Ramanath, R., Liu, F., Sadeh, N., Smith, N.A.: Unsupervised alignment of privacy policies using hidden markov models. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 605–610 (2014)
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. Sathyendra, K.M., Schaub, F., Wilson, S., Sadeh, N.: Automatic extraction of opt-out choices from privacy policies. In: 2016 AAAI Fall Symposium Series (2016)
19. Sathyendra, K.M., Wilson, S., Schaub, F., Zimmeck, S., Sadeh, N.: Identifying the provision of choices in privacy policy text. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2774–2779 (2017)
20. 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 (2015)
21. Shayegh, P., Ghanavati, S.: Toward an approach to privacy notices in iot. In: 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW). pp. 104–110. IEEE (2017)
22. Shayegh, P., Ghanavati, S.: Toward an approach to privacy notices in iot. In: 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW). pp. 104–110. IEEE (2017)
23. Shayegh, P., Jain, V., Rabinia, A., Ghanavati, S.: Automated approach to improve iot privacy policies. arXiv preprint arXiv:1910.04133 (2019)
24. Siboni, S., Shabtai, A., Tippenhauer, N.O., Lee, J., Elovici, Y.: Advanced security testbed framework for wearable iot devices. ACM Transactions on Internet Technology (TOIT) **16**(4), 1–25 (2016)
25. Singh, R.I., Sumeeth, M., Miller, J.: Evaluating the readability of privacy policies in mobile environments. International Journal of Mobile Human Computer Interaction (IJMHCI) **3**(1), 55–78 (2011)
26. Statista Research Department: Internet of things (iot) connected devices installed base worldwide from 2015 to 2025.

https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ (2020)
27. Subahi, A., Theodorakopoulos, G.: Ensuring compliance of iot devices with their privacy policy agreement. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud). pp. 100–107. IEEE (2018)
28. Sunyaev, A., Dehling, T., Taylor, P.L., Mandl, K.D.: Availability and quality of mobile health app privacy policies. Journal of the American Medical Informatics Association **22**(e1), e28–e33 (2015)
29. 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: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1330–1340 (2016)

A Appendix A - IoT manufacturers

List	IoT Manufacturer	The IoT manufacturer PPA URL
1	TP-link	https://www.tp-link.com/uk/about-us/privacy/
2	Belkin netcam	https://www.belkin.com/us/privacypolicy/
3	Lifx	https://www.lifx.com/pages/privacy-policy/
4	Hive	https://www.hivehome.com/privacy
5	Philips hue	https://www2.meethue.com/en-gb/support/privacy-notice
6	Awair	https://getawair.com/pages/legal#privacy
7	Smart Things	http://www.smartthings.com/gb/privacy
8	Nest	http://nest.com/uk/legal/privacy-statement-for-nest-products-and-services/
9	Elgato Avea	https://www.elgato.com/en/data-protection
10	Ikea Tradfri	https://www.ikea.com/gb/en/customer-service/privacy-policy/
11	Eufy Lumos	https://www.eufylife.com/uk/privacy-policy
12	Nanoleaf	https://nanoleaf.me/en/privacy/
13	Osram Lightify	https://www.osram.com/cb/services/privacy-policy/index.jsp
14	Sengled Element	https://eu.sengled.com/en/about-us/privacy-policy/index.html
15	Xiaomi	https://privacy.mi.com/all/en_GB/
16	LOHAS	https://www.lohas-led.com/art/privacy-policy-a0040.html
17	Devolo	https://www.devolo.co.uk/support/data-privacy
18	Arlo	https://www.arlo.com/en-us/about/privacy-policy/
19	Ring	https://en-uk.ring.com/pages/privacy-notice
20	Swann	https://www.swann.com/uk/company/privacy-policy
21	D-Link	https://eu.dlink.com/uk/en/privacy
22	Neos	https://shop.neos.co.uk/pages/privacy-policy
23	Logi	https://www.logitech.com/en-gb/legal/web-privacy-policy.html
24	Ezviz	https://www.ezvizlife.com/uk/legal/privacy-policy
25	Netatmo	https://view.netatmo.com/uk/legals/app?gsc=true&goto=privacy
26	Blink XT	https://blinkforhome.co.uk/pages/privacy-policy
27	Canary	https://canary.is/legal/privacy-policy/
28	Somfy	https://www.somfy.co.uk/privacy-policy
29	Samsung	https://www.samsung.com/uk/info/privacy/
30	Google Home	https://policies.google.com/privacy
31	Toymail	https://toymail.co/pages/privacy
32	Resideo	https://www.resideo.com/us/en/corporate/legal/privacy/english/
33	Alexa	https://www.alexa.com/help/privacy
34	Fossil	https://support.fossil.com/hc/en-gb/articles/360026199151-WEBSITE-PRIVACY-AND-SECURITY
35	Mobvoi	https://www.mobvoi.com/us/pages/privacy-policy
36	Fitbit	https://www.fitbit.com/uk/legal/privacy-policy
37	Brilliant	https://www.brilliant.tech/pages/privacy-statement-for-brilliant-products-and-services
38	Wink	https://www.wink.com/legal/
39	Wyze	https://wyze.com/privacy-statement-wyze-site
40	August	https://august.com/pages/privacy-policy#product
41	SimpliSafe	https://simplisafe.co.uk/privacy
42	Ecobee	https://www.ecobee.com/legal/use/
43	Anova	https://anovaculinary.com/privacy/
44	Ecovacs	https://www.ecovacs.com/global/company/common?type=privacypolicy
45	iRobot	https://www.irobot.com/legal/privacy-policy
46	Nanit Plus	https://www.nanit.com/uk/legal/privacy
47	Sleep Number	https://www.sleepnumber.com/legal-notices/privacy-policy
48	Fluidrausa Smart Outdoor Gadgets	https://www.fluidrausa.com/en/legal
49	Rachio	https://www.rachio.com/privacy-policy/
50	Traeger Ironwood	https://www.traegergrills.com/privacy-policy

Table 4. IoT manufacturers PPA URL