"PrivacyLens: A Framework to Analyze the Landscape of Past, Present, and Future Smart Device Privacy Policies using Machine Learning"

Abstract

In this research paper, we present a comprehensive and automated framework for the analysis and evaluation of privacy policies associated with Internet of Things (IoT) devices. Our framework utilizes machine learning algorithms for feature extraction and ambiguity detection, while leveraging web data and historical information to assess privacy policies across a range of devices.

We demonstrate the effectiveness of our proposed framework by applying it to a representative sample of IoT devices. Our results indicate that the framework can accurately identify critical components of privacy policies and uncover potential ambiguities that may impact the privacy quality of IoT devices. Additionally, we compare our findings with the Mozilla Privacy Not Included database, revealing that our framework offers comparable insights into the privacy quality of IoT devices.

The primary objective of our research is to address the challenges and limitations inherent in current IoT privacy policy practices. By providing valuable insights, our work aims to enhance the usability and effectiveness of these policies for end-users. Through the development of a robust framework for evaluating and contrasting privacy policies, as well as understanding their design, we strive to safeguard user data and privacy in the rapidly expanding IoT ecosystem. Our research plays a crucial role in fostering increased transparency and accountability in the digital era, ultimately empowering users to make better-informed decisions when using digital services, thereby protecting their personal data and privacy.

1 Introduction

The Internet of Things (IoT) has rapidly gained popularity in recent years, with smart devices being installed at a rapid pace in various domains such as transportation, industrial processes, personal smart homes, and health care. Alter et al. [1] defined smartness, in the context of smart devices and systems, as a collection of continuous variables or dimensions that vary from not at all smart to highly smart. More than 40 million households in the US have already adopted smart home devices, and this number is expected to quadruple by 2022 [2]. However, the growing use of smart technology also poses risks to privacy and security [3].

IoT devices collect large amounts of diverse data, and consumers are sometimes even unaware that data is being collected in their environment. As a result, it is unclear what type of data is collected, used, and transferred by smart devices. Data sharing may allow companies to create complete profiles of their users or allow attackers to disrupt normal behaviour [4]. The potential privacy implications of IoT devices have raised concerns among consumers, particularly because smart devices collect a lot of personal information about them. Most businesses have privacy policies that outline the types of information collected as well as how it is used, disclosed, and maintained. Privacy policies have traditionally provided information about the data collected by websites, and extensive analysis has been performed on them. Attempts have been made to assess the readability of website privacy policies and demonstrate how they can influence user behaviour [5, 6]. Miller et al. [7] developed a metric for assessing the completeness and content of web privacy policies, client storage and tracking practices, data handling practices and policies, and usability. However, it is not clear whether smart devices have accessible privacy policies and whether they are good enough, as not much analysis has been conducted on them.

In this paper, we present a framework for analyzing the privacy policies of various smart devices. We created a curated corpus of 462 smart device privacy policies and analyzed them using various metrics such as similarity, ambiguity, key information, keyword analysis, and named entities, and then compared them. Our goal is to understand the current state of privacy policies for smart devices and to identify areas for improvement in terms of transparency and user-friendliness.

To perform our analysis, we first needed a corpus of privacy policies from a diverse set of smart devices. We selected a sample of devices from different categories such as smart home devices, wearables, and industrial devices. Our frame-

work allowed us to collect policies quickly and efficiently, while still ensuring that we had a representative sample. We used various metrics to analyze the policies, such as the similarity of the policies, the level of ambiguity in the language used, and the presence of key information such as user access, control, and deletion. Our framework enabled us to perform these analyses quickly and accurately, providing us with valuable insights into the current state of privacy policies for smart devices.

For the similarity metric analysis, we calculated the cosine similarity scores of the IoT device privacy policies and grouped them based on similarity scores, yielding outliers. For the ambiguity analysis, we manually labelled 100 privacy policies into three categories: 'somewhat ambiguous,' 'not ambiguous,' and 'very ambiguous.' We then used the labelled dataset to train a random forest classifier, which classified the dataset into the three categories listed above. We also performed key information analysis to identify which product categories were mentioned in the privacy policies and presented the results using bar graphs.

Our analysis provides insights into the current state of privacy policies for smart devices. We found that there is a lack of consistency and clarity in the way that privacy policies are presented across different smart device manufacturers. Our findings suggest that smart device users may not fully understand the types of data that are being collected and shared by their devices, which could pose a significant threat to their privacy. To address these issues, we provide recommendations for improving the clarity and transparency of smart device privacy policies. Our recommendations include using clear and concise language, providing examples of data handling practices, and creating more accessible formats for privacy policies. We also recommend that policymakers consider regulating the privacy policies of smart device manufacturers to ensure that they provide adequate protection for consumers.

Overall, our automated framework allowed us to collect and analyze a large corpus of privacy policies for smart devices. This enabled us to identify trends and patterns in the policies, as well as areas where policies could be improved. By automating the process of collecting and analyzing policies, we were able to perform our research efficiently and effectively, providing valuable insights into the state of privacy policies for smart devices. It highlights the need for more accessible and user-friendly policies that clearly outline data handling practices and protect users' privacy. Our framework can be used to evaluate and improve privacy policies for smart devices, ensuring that users can make informed decisions about their data privacy.

The remainder of this paper is structured as follows. In Section 2, we review related literature on privacy policies in IoT devices and discuss the challenges and limitations of these policies. In Section 3, we describe our research methodology, including the corpus of privacy policies, the metrics and methods used in the analysis, and the human annotators who verified the results. In Section 4, we present the results of our analysis, including the similarity, ambiguity, key information, keywords, and named entities in the privacy policies. In Section 5, we discuss the implications and limitations of our research and provide recommendations for future work. Finally, in Section 6, we conclude the paper with a summary of our findings and contributions.

2 Related Work

Several methodologies have been developed for capturing individual privacy policies of IoT devices [8]. In this section, we review related work on annotated datasets, automated frameworks, ambiguity detection, and text similarity in the context of privacy policy analysis for IoT devices.

2.1 Annotated Datasets

Annotated datasets have been widely used for training machine learning models to analyze privacy policies. The OPP-115 dataset [**?**] is a corpus of 115 website privacy policies that were collected using Amazon Alexa's technology [**?**]. The dataset includes annotations and a labelling system created by its authors, which provides examples of how personal data is used and details on the experts who annotated the texts. Another annotated dataset is the APP-350 corpus, which includes over a million Android application privacy policies available on the Google Play store [**?**]. However, these annotations were created prior to the adoption of the GDPR, and therefore do not take into consideration the requirements of this regulation.

2.2 Automated Frameworks

Several automated frameworks have been proposed for analyzing and evaluating privacy policies for IoT devices. The first general framework that allows for comprehensive automated examination of privacy regulations was introduced to help consumers, researchers, and regulators process and comprehend privacy rules on a large scale [9]. Polisis, a framework developed by the same authors, uses a neural network hierarchy to extract high-level privacy practices and precise data from privacy regulations [9]. This allows for both structured and free-form querying of privacy policies and has been used to create applications such as privacy icons and PriBot.

Another automated framework was proposed by Kuznetsov et al. [8], which implemented a novel approach for gathering privacy policies from IoT devices. Unlike previous methods that relied on the Amazon Alexa service, this approach starts with e-commerce sites to generate a document corpus. The researchers used this program to collect their own datasets, which included 592 distinct privacy policies from various IoT device manufacturers.

2.3 Ambiguity Detection

Ambiguity in privacy policies can contribute to a lack of understanding and hinder effective analysis. To address this, researchers have developed methods for ambiguity detection in privacy policies. Kotal et al. [10] presented a mechanism for categorizing policy papers based on their ambiguity and extracting factors from policy statements that influence ambiguity. The researchers validated the strategy using human annotators and showed that a substantial part of the documents in a well-known corpus of privacy policies (OPP-115) is ambiguous.

2.4 Text Similarity

Text similarity is important for understanding how people comprehend language, and several studies have investigated various measures of text similarity for privacy policy analysis. One study by Resnik [11] proposed using the degree of informational overlap between two ideas to indicate their similarity, and derived a semantic similarity measure based on this. Another study by Nwachukwu et al. [12] analyzed the keywords and content of over 2000 online policies and used topic modelling algorithms to analyze topic coverage and measure the coverage of ambiguous words in privacy policies.

In conclusion, the reviewed studies demonstrate a range of approaches for analyzing and evaluating privacy policies for IoT devices. Annotated datasets, automated frameworks, ambiguity detection, and text similarity are all important areas of research that can provide valuable insights into privacy quality and promote greater transparency and accountability in the digital age.

Annotated datasets, such as OPP-115 and APP-350, have been used to train machine learning algorithms for privacy policy analysis. These datasets provide labelled examples of different uses of personal data in privacy policies and enable researchers to evaluate the effectiveness of their methods. However, these datasets were created prior to the adoption of GDPR and may not reflect the current legal requirements for privacy protection.

Automated frameworks, such as Polisis and the framework proposed by Kuznetsov et al., provide a comprehensive approach for analyzing and evaluating privacy policies at scale. These frameworks use machine learning algorithms to extract key information from privacy policies and generate interpretable insights into privacy quality. They also have the potential to support compliance assessments with legal and regulatory frameworks, such as GDPR and CCPA.

Ambiguity detection is an important area of research that addresses the challenge of interpreting natural language privacy policies. Strategies for categorizing policy papers based on their ambiguity and extracting factors from policy statements that influence ambiguity have been proposed, and these approaches can provide insights into areas of potential confusion or misunderstanding in privacy policies.

Text similarity measures have been used to compare and evaluate the similarity of different privacy policies. These measures can provide insights into common themes and practices across different privacy policies and enable researchers to identify areas of concern or best practices.

Overall, the reviewed studies highlight the importance of analyzing and evaluating privacy policies for IoT devices to ensure greater privacy protection for users. Future research can build upon these approaches to further improve the accuracy and effectiveness of privacy policy analysis and address the ongoing challenges associated with privacy protection in the digital age.

3 Methodology

In this part, we'll go into great depth on how we selected a broad range of privacy policies of Iot devices, our annotation system, how we gathered annotations, what criteria we used to choose the policies and the structure of the chosen corpus [13].

3.1 Privacy Policy Selection

A privacy policy describes each procedure a company uses to gather, handle, and distribute user data. While some contend that privacy rules serve solely as informational documents [14], others maintain that they are binding contracts that must be followed [15]. Although privacy policies frequently contain less legalese and are easier to comprehend than terms and conditions, terms of service, or end-user license agreements, they nevertheless contain a lot of the same information. Good privacy policies are forthright, understandable, and transparent. The selection of the privacy policies was based on the following questions:

1. Can consumers make informed decisions about IoT devices before they consider buying them?
2. Are IoT device privacy policies clear to consumers?
3. Do smart device privacy policies differ significantly from each other?
4. Do we see any patterns regarding the effectiveness of IoT rules (e.g., regarding the manufacturer's country of origin or the type of device?

These questions were important because they provide key insights from the policy document about user control and choices. The initial stage in creating our data corpus was to acquire a predefined dataset [8] so that we could focus on analysis. We picked this dataset since it was consistent with our study concept and emphasized the collection of privacy policies produced by manufacturers of smart IoT devices. The authors created the corpus by breaking down the process

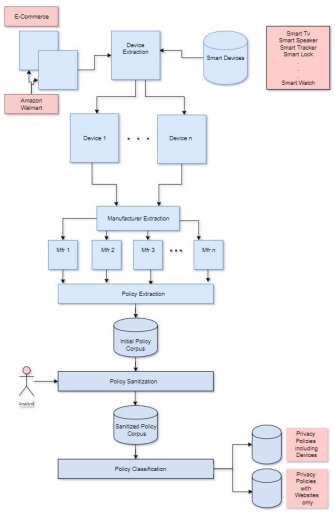


Figure 1: Steps to obtain the IoT privacy policy corpus.

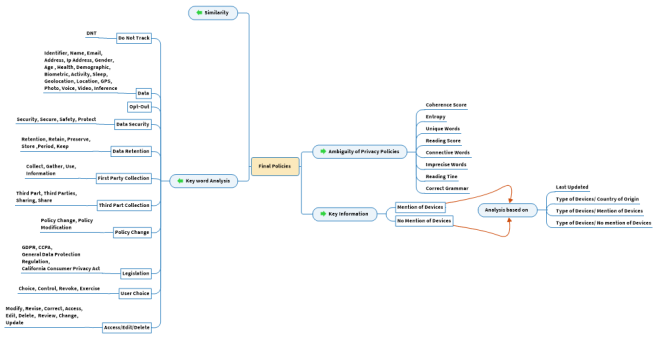


Figure 2: Steps to obtain the features from the corpus.

of collecting privacy policies into smaller tasks. The user's first-phase selection of e-commerce platforms is used by the crawler [16] to gather connections to the websites of IoT device manufacturers. The user may define more than only IoT devices because the device type is an input parameter. Then the page markup is used to obtain the manufacturer's information. The next step is to do an online search for manufacturers of smart devices. If the manufacturer is found, the web page looks for a link to the privacy policy. The URLs to the privacy policies are downloaded when collected [17]. After that, HTML markup and code are removed from the privacy rule's content to provide structured information in simple English. Next, the authors collected a corpus of 803 privacy policies using the above-mentioned method. Then we manually checked all the policies in the corpus to ensure they were similar concerning IoT devices. After an extensive investigation, we found two critical flaws in the corpus. Firstly, 401 policies were not representing IoT devices at all. These policies belonged to websites of news agencies, online marketplaces, and companies that shut down long ago. Another critical flaw in the corpus was that the popular IoT device's policies were missing. To come up with a standard corpus that was better suited for the study, we added another 60 policies [18], [19] which consisted of 50 policies for smart IoT devices and ten policies for smart car's as they also act as smart devices. The final data set consisted of 462 policies.

3.2 Annotation Scheme and Process

This study deals with two types of annotation schemes. First, to capture the data practices described by privacy policies, we employed a policy annotation approach [13]. The eight data practice categories that make up the final annotation for this category are:

1. First Party Collection/Use: The methods and purposes used by a service provider to get user data.
2. Third Party Sharing/Collection: The methods used by third parties to share or acquire user information.
3. User Choice/Control: The alternatives that users have for choice and control.
4. User Access, Edit, and Deletion: if users may access, edit, or remove their information, and how.
5. Data Retention: How long is user data kept on file?
6. Data Security: How user data is safeguarded.
7. 7 Policy Change: Whether and how users will be informed of privacy policy changes.
8. 8 Do Not Track: Whether and how internet tracking and advertising using Do Not Track signals are handled.

Second one, is to capture the key information and observable characteristics about the policies [10]. An important contribution to our study is to determine the level of ambiguity in a policy by using important characteristics of natural language. There are numerous methods for detecting ambiguity in Natural Language [20–23].Different linguistic characteristics of a text that influence its ambiguity have been defined in previous research on other technical publications [24–26]. This section describes eight quantifiable characteristics of a policy text that contribute to its ambiguity. We outline these characteristics and show how they may be extracted for use in a policy text [10].

1. Frequency of Imprecise Words: Some terms are naturally imprecise in the English language. Texts may become unclear if such terms are used frequently. For instance, it may be difficult to understand general terms like "commonly" or "normally," which give an incorrect picture of the service provider's operations and cause uncertainty in determining the statement's true meaning.

Imprecise Words	
Modal Words	may,might,likely,can could,would
Usable Words	easy,adaptable familiar,extensible
Probable Words	probably,possibly,optionally
Numeric Words	anyone, certain everyone, numerous some,most,few much,many,various including but not limited to
Condition Words	depending,necessary inappropriate,appropriate as needed,as applicable otherwise reasonably from time to time
Generalization Words	generally,mostly,widely commonly,usually,general Normally,typically,largely often,primarily among other things

Table 1: Taxonomy of Imprecise Words

Connective Words	
Copulative Words	and, both, as well as, not only, but also
Control Flow Words	if, then, while
Anaphorical Words	it, this, those

Table 2: Taxonomy of Connective Words

2. Frequency of Connective Words: In the English language, connecting words are used to join clauses or phrases. They are crucial for the creation of coherent sentences. However, overusing connectives makes a text more complicated. The following is an excerpt from a Policy that might be interesting to observe: "Like most online service providers, we collect information that online services, portable devices, and cloud services typically make available, including the internet browser, Source IP, unique device identifiers, language choice, referring site, the date and time of access, operating system, and mobile network information". Three times in this phrase, the conjunction "and" has been used to connect different clauses and reading this sentence is challenging. Research related to textual information [**?**], [**?**] suggests counting the number of connective terms to assess the text's quality in software requirements. In Table 3.2 [10] have developed a taxonomy of connective terms which we use in our framework. We counted how often these connecting terms appear in Policies. This metric helped us determine how ambiguous a document was overall.

3. Readability Score: Readability refers to how simple or difficult it is to read something. The readability of a text is determined by how it is presented and the context in which words and phrases in the document are delivered. Additional elements that influence reading are sentence length, sentence structure, and syllable count per word. These aspects work together to determine how well your writing will be comprehended. Readability is crucial since it determines how well a reader understands a piece of text. Policies should be easy to read because they describe the policies and practices of the businesses regarding the data that is gathered from consumers. There are many readability tests [27] which have been devised by linguists and are based on different factors like the vocabulary used, the syntax and the sentence structure. For our study, we used Flesch-Kincaid Grade Level [28] which displays the score in terms of a U.S. grade level. It demonstrates the level of education needed to comprehend a text given by formulae:

$$0.39\left(\frac{totalwords}{totalsentences}\right)+11.8\left(\frac{totalsyllables}{totalwords}\right)-15.59 \quad (1)$$

We used the readability score that the Flesh-Kincaid grade level method assigned the document in our approach. This is one of the metrics we employed to gauge how ambiguous a policy paper was.

4. Correct Grammar: The integrity of a work depends on proper grammar, much as it does on word spelling. Every sentence of a policy document went through the framework [29] to ensure that the grammar was used correctly. The Python Language-Check package was employed for this. Each tokenized text which was produced with the help of NLP is compared to a parse tree to ensure that the grammar is valid. We track the proportion of sentences with poor grammar in each text relative to the overall number of sentences.

5. Entropy: A statistical quantity called the entropy of a language gauges, in a sense, how much information is generated on average for each letter in a text in that language. In simpler terms, it defines the uncertainty or disorder in a text document. Shannon [30] presented a novel technique of calculating the entropy of English text arguing that everybody who speaks the language is well knowledgeable of the language's characteristics. The uncertainty of the meaning of textual terms is expressed in two ways: the meaning we get from it and the context in which it is utilized. The equation for generating the entropy is given by:

$$H(X)=H(P_1,...,P_n=-\sum_{i=1}^{n}P_iLog_2P_i \quad (2)$$

Entropy was computed by iterating over the policy document's characters and noting the frequency of occurrence of each one. We divided the frequency by the total

number of characters to get an estimate of each character's probability. Each character's average length in bits was then calculated by multiplying its likelihood by the negative logarithm of that same probability. Finally, we summed the average lengths of all characters to get the result.

6. Reading Time: According to study [31], privacy policies are hard to read, rarely read, and don't help customers make informed decisions due to the fact that they are very lengthy and time-consuming. Reading time is calculated using an individual's typical reading pace (roughly 238 WPM) [**?**]. We used an approach that involved counting all of the words in the document and dividing the total by 238 to produce a decimal number. The minute is the first digit of the decimal number. Then we multiply the second half (the decimal points) by 0.60, these are the seconds. We round up to get a whole number that represents the total reading time of the policy.

7. Frequency of unique words: The goal of the privacy policy document is to provide consumers with relevant and crucial information about how their data is used and if they have control over that data. This knowledge is conveyed using specialized technical jargon which can sometimes be hard to comprehend based on the context. The bounds of certain learning categories are being pushed by these unusual words, and the words around them offer crucial context. People usually don't encounter low-frequency words which makes it hard for them to understand and comprehend these works which leads to the argument that unique words require unique instruction [32], because informational texts integrate vocabulary and concepts, it is hard to completely comprehend the material without a solid understanding of the unique, low-frequency phrases. It is hard to skip through complicated jargon while still having a general understanding of the subject matter; therefore, individuals must have expert-level knowledge of these words in order to grasp the content. A relatively narrow vocabulary that makes up 90% of the words in texts includes the most commonly used terms. This word set includes a foundation of 4,000 simple word families and produces around 10,000 words [33]. So unique words in documents are key factors in gauging the effectiveness of the policy text. For this metric, we used NLP [34] to first clean the text from stop words, punctuation's and numbers and then tokenized the text. After that, we extracted unique words in each document with Spacy [35]. By dividing the number of unique words by the total number of words in the document, we were able to calculate the frequency of unique words.

8. Coherence Score: How easily the topics are understood by people can be determined using the coherence score in topic modeling [36]. The top N words that are most likely to fall under a certain topic are displayed as topics. The general similarity of these words to one another is measured by the coherence score. It identifies the likelihood that a word or phrase belongs to a particular topic and group topics together depending on how similar or dissimilar they are. It accomplishes this by looking at how often certain words and phrases appear in the document. Latent Dirichlet Allocation [37] is a machine learning, unsupervised clustering method that is employed for text analysis. We proceed with the assumption that the policy document is made of several topics with each topic containing different words and the LDA topic modelling algorithm model is capable of inferring the topics that better represent the data. For this analysis we use LDA to first create a Dirichlet distribution of documents in the subject space, then choose N topics from a multinomial distribution of subjects over a document. The second stage starts by choosing N words from the multinomial distribution of words over subjects for each of the previously selected topics and begins the Dirichlet distribution of topics in the word space. Finally increasing the likelihood that the same topics will be produced. Once we generate N topics we generate the coherence score as:-

$$Coherence = \sum_{i<j} \text{score}(w_i, w_j) \tag{3}$$

The coherence score is for assessing the quality of the learned topics and the score better the representation. We are using One of the most used coherence measures CV which constructs content vectors for the words using word co-occurrences and then computes the score using cosine similarity and normalized pointwise mutual information (NPMI).

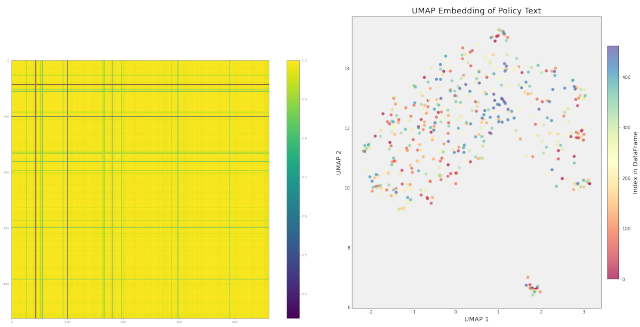
4 Analysis of Results

In this chapter, we present the results of our study. We conducted four fundamental analyses, and we present our findings based on those analyses. The results are divided into four different subsections.

4.1 Policy Similarity

Text similarity is the process of comparing a piece of text with another and finding the similarity between them. It is a part of natural language processing (NLP), which involves processing raw text to detect similarity. To start with our text similarity task, we used the Word2Vec algorithm, which uses neural networks to extract the contextual meaning of the corpus and develop the embeddings using the continuous bag of words (CBOW) method [38].

The CBOW model is a supervised learning neural network model that predicts the center word from the corpus words. It



(a) Heatmap of Policy Embeddings (b) PCA of Policy Embeddings

Figure 3: Visualizations of Policy Embeddings

takes one-hot encoding of the words as input, and the output is the main word that can add some sense to the neighbouring terms. The weights in the hidden layers of the Word2Vec model are the embeddings we need. Word embedding is one of the most popular representations of document vocabulary. It can capture the context of a word in a document, semantic and syntactic similarity, relation with other terms, etc.

Finally, we measure the similarity between two vectors of an inner product space. It calculates the cosine of the angle between two embeddings and determines whether they are pointing in roughly the same direction or not. When the embeddings point in the same order, the angle between them is zero, so their cosine similarity is 1. When the embeddings are perpendicular, the angle between them is 90 degrees, and the cosine similarity is 0. When the angle between them is 180 degrees, the cosine similarity is -1. The similarity between the two documents is given by:

$$Similarity(X,Y) = \frac{X.Y}{|X| \times |Y|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (4)$$

where X and Y are the two embeddings to be compared, and n is the dimensionality of the embedding space. Several studies have used the cosine similarity measure for policy similarity analysis, such as [3, 4, 26].

In this study, we used conventional methods to establish a baseline for measuring document similarity in the policy domain. The objective was to gain insights into how well these metrics perform in this domain. The results indicate that nearly all policies (97%) are similar, with little variation in their metric values, rendering them identical based on the chosen technique. However, we identified a few outliers (3%) that differed significantly from the others. This finding motivated us to develop a new set of metrics that not only measure similarity but also aid in the analysis of ambiguity in policies.

Table 2 represents the policy features of different companies. These features include Coherence Score, Entropy, Freq. of Unique Words, Reading Complexity, Reading Time, Freq. of Imprecise Words, Freq. of Connective Words, and Correct Grammar.

Company Name	Coherance Score	Entropy	Unique Words	Reading Time	Reading Level	Imprecise Words	Connective Words	Correct Grammer
Eco4lifehome	0.72	10.29	0.27	7	15.10	0.01	0.01	0.52
Nooie	0.70	5.49	0.67	9	4.59	0.02	0.03	0.31
Fdt	0.92	6.86	0.34	3	5.89	0.01	0.02	0.15
Axis	0.42	5.41	0.54	4	5.34	0.01	0.01	0.14
Mobvoi	0.25	7.26	0.45	6	4.24	0.01	0.04	0.18
Alarmlock	0.81	7.14	0.52	2	5.76	0.01	0.01	0.27
Umidigi	0.59	6.96	0.33	2	4.69	0.01	0.02	0.1
Evapolar	0.22	8.55	0.56	6	13.39	0.01	0.05	0.32
Cablematters	0.74	7.01	0.37	3	6.43	0.01	0.05	0.31
Bulbrite	0.80	7.94	0.40	5	5.96	0.02	0.04	0.52
Airivo	0.61	7.88	0.50	3	6.16	0.01	0.04	0.26
Luxproducts	0.55	7.41	0.44	2	4.38	0.01	0.04	0.19
Adero	0.58	7.56	0.45	2	4.77	0.01	0.05	0.21

Table 3: Policy Features

Policy Features	Min Value	Average Value	Max Value
Coherence Score	0.13	0.35	0.92
Freq. of Imprecise Words	0	0.02	0.2
Freq. of Connective Words	0.01	0.04	0.08
Reading Complexity	4.24	11.40	21.73
Reading Time (Min)	2	12.67	107
Entropy	5.41	7.97	10.29
Freq. of Unique Words	0.10	0.30	0.67
Correct Grammar	0	0.25	1.06

Table 4: Min, Average and Max Feature values extracted from corpus

Upon analyzing the results, it can be observed that the companies have different scores for each of the policy features. The coherence scores of the companies range from 0.22 to 0.92, while the entropy values range from 5.41 to 10.29. The Freq. of Unique Words ranges from 0.27 to 0.67, and the reading complexity scores range from 4.24 to 21.73.

The reading time scores of the policies range from 2 to 107 minutes, with some policies having a relatively lower reading time. The frequency of imprecise words ranges from 0.01 to 0.02, while the frequency of connective words ranges from 0.01 to 0.05. Finally, the correct grammar scores range from 0.1 to 0.52.

It can be observed that there is no particular company that has the best or worst score in all the policy features. Each company has a unique set of scores for different policy features. These scores indicate how clear, concise, and informative the policies of these companies are. The companies with higher scores for coherence, reading complexity, and correct grammar can be assumed to have better policies that are easy to read and understand. On the other hand, companies with lower scores for these features may have policies that are less clear and more difficult to understand.

We analyzed the similarity metrics of the policies using conventional methods as a baseline approach. The results revealed that 97% of the policies were similar and lacked any significant difference between them as they had similar feature values, which made them identical according to the technique. However, the remaining 3% of the policies formed the outlier group, consisting of 13 policies with different feature values than their counterparts. Figure 4.1 shows a heat map representation of the outlier group, where the dark lines represent the policies that are most dissimilar to other policies.

To explain the heat map, the darker the area, the warmer

the shades of yellows, indicating the most dissimilar policies, while areas with a similarity between 0.5 and 0.97 on average with other policies are marked in green colours. The eco4homelife policy was the most dissimilar to other policies, while Adero had the highest average similarity across the group. Further investigation showed that the outlier policies had lower-than-average reading time and reading complexity.

We found a correlation between the dissimilarity of policies and higher-than-average average scores across the metrics described. Word embeddings exist in very high dimensionality, making it impossible to visualize how words occupy the embedding space. Hence, we used the Principal Component Analysis (PCA) method to reduce the dimensionality of word embeddings and visualized them in a 2D space, as shown in Figure 4.2. The graph gave similar indications that there were three groups: one with an average similarity score of 0 to 92% with other policies, the second with an average score greater than 92% and less than 97%, and the third group whose average similarity was more significant than others, i.e., greater than 97% on average.

It was further revealed that the policies in the outlier list had higher-than-average entropy scores, higher-than-coherence scores, and relatively lower-than-average reading levels and reading times, making them less challenging to read and producing more engaging sentences but lacking vital information that should have been present in the policies. Additionally, policies in the outlier group had a lower-than-average frequency of connective words and imprecise words.

4.2 Ambiguity Analysis

In this analysis, we selected eight metrics that represented the policies regarding key information. To classify the policies based on these metrics, we used a supervised learning algorithm known as the random forest classifier [39].

Supervised learning is a reliable approach for classification tasks but requires a significant portion of the dataset to be labelled. Therefore, we partially annotated 100 policies in our dataset, which was sufficient for the random forest classifier to work well for our dataset.

Random forest is a classification technique that uses a group of classification trees. Each tree is built using a bootstrap sample of the data and a random subset of variables for each split. This approach uses bagging, a method for aggregating unstable learners, and random variable selection to form trees. To produce low-bias trees, each tree is left unpruned, but the combination of bagging and random variable selection leads to little correlation among the individual trees. By averaging across a considerable ensemble of low-bias, low-variance data, the technique produces an ensemble that can achieve low bias and low variance.

The objective of the random forest classifier is to classify policies based on a training set T = $(t_1,...,t_n)$ with responses R = $(r_1,...,r_n)$. Bagging is used to repeatedly select a random sample with the replacement of the training set and fit trees to these samples such that $n = (1,...,N)$. After training, the majority votes are taken to classify a new instance.

Ambiguity Class	Number of Policies	Random Forrest Classifier	Logistic Regression
Not Ambiguous	283	0.86	0.77
Somewhat Ambiguous	90	0.67	0.72
Very Ambiguous	89	0.67	0.69

Table 5: F1-score of Random Forrest Classifier for each Ambiguity Class

Manufacturer Country	Not Ambiguous	Very Ambiguous	Somewhat Ambiguous
USA	60.8%	18.7%	27.5%
China	54.3%	23.9%	21.7%
European Union	65%	22.5%	12.5%

Table 6: Percentage of Ambiguity Level across Manufacturers

The F1 score was used as a measure of accuracy, which is the harmonic mean of precision and recall [40]. The F1 score is given by Equation (1):

$$F1\ Score = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)} \tag{5}$$

Here, recall is the fraction of relevant instances that were correctly classified, and precision is the fraction of instances classified as positive that were correct. The F1 score ranges from 0 to 1, with a higher score indicating better accuracy.

Our results showed that the classification of policy texts is significantly more challenging for more ambiguous policies. As seen in Table 4.4, we observed a drop in the F1 score from 86% to 67% as the ambiguity of the policy increased. Additionally, in the case of logistic regression, we observed that the F1 score dropped from 77% to 69% when moving from non-ambiguous to ambiguous policies. These findings indicate that classification algorithms have less accurate results for highly ambiguous policies.

To further analyze the performance of classification models, we evaluated their accuracy in each ambiguity category of documents. Our findings showed that the classification algorithms had less accurate results for highly ambiguous policies. This could be attributed to the fact that ambiguous policies have less well-defined requirements and guidelines, making it harder for classification algorithms to accurately identify them. Also, our results indicate that policy ambiguity is a significant factor that affects the accuracy of classification models, and more research is needed to develop algorithms that can accurately classify ambiguous policies.

Our analysis revealed interesting insights into the level of ambiguity in privacy policies across different manufacturers and regions. Firstly, we found that the policies in the outlier group, which had different feature values than their counterparts, were mostly classified as non-ambiguous by the classification algorithm, with the exception of one policy from "Bulbrite" which was classified as very-ambiguous, indicating a possible misclassification by the algorithm.

Additionally, we observed that almost all the policies of smart-connected cars were classified as very-ambiguous, sug-

gesting that these policies need to be improved to provide clear and concise information to users. We also analyzed the ambiguity levels of policies across different regions and found that EU privacy policies tend to have a lower ambiguity level than their counterparts from the US and China. This could be attributed to the stricter data privacy laws and GDPR in the EU, which compel companies to provide more detailed information to users.

Furthermore, we categorized the policies into three levels of ambiguity - not-ambiguous, somewhat-ambiguous, and very-ambiguous - and found that EU had the highest percentage (65.8%) of policies falling in the not-ambiguous category, while China had the lowest (54.3%). In the somewhat-ambiguous category, the US topped the list with (27.5%) policies and EU had the lowest percentage (12.5%). In the very-ambiguous category, China had the highest percentage (23.9%) and the US had the lowest (18.7%).

Finally, analysis of the data revealed an interesting insight, which is that policies that did not mention devices had a higher percentage of policies in the not-ambiguous category and a lower percentage in the very-ambiguous category when compared to policies that mentioned devices, as shown in Figure 4.3. We interpret this finding as suggesting that policies that do not mention devices tend to be less ambiguous overall because they focus more on general data handling and user control practices, rather than specific technical details. This observation is consistent with previous research, which has shown that policies that focus more on high-level privacy principles tend to be more accessible and understandable to users [41].

Overall, the results suggest that companies need to pay more attention to the level of ambiguity in their privacy policies, particularly for smart-connected cars. The findings also highlight the need for stricter regulations and guidelines to ensure that privacy policies are more transparent and easier to understand for users.

4.3 Key Information Analysis

Data privacy regulations have become a crucial part of the online world, with many countries around the globe implementing their own laws and regulations to safeguard users' private information. Websites or devices that collect personal data must comply with these regulations, and failure to do so can result in severe consequences such as heavy fines and even prosecution. Leading organizations have adapted to these regulations, and compliance has become the norm.

In our analysis, we aimed to investigate whether privacy policies from different countries followed a different approach to collecting, using, and storing user data, and if additional data privacy regulations had an impact on privacy policies originating from different countries. To achieve this, we manually collected the origin of IoT devices by carefully checking the countries where the manufacturers were based for each

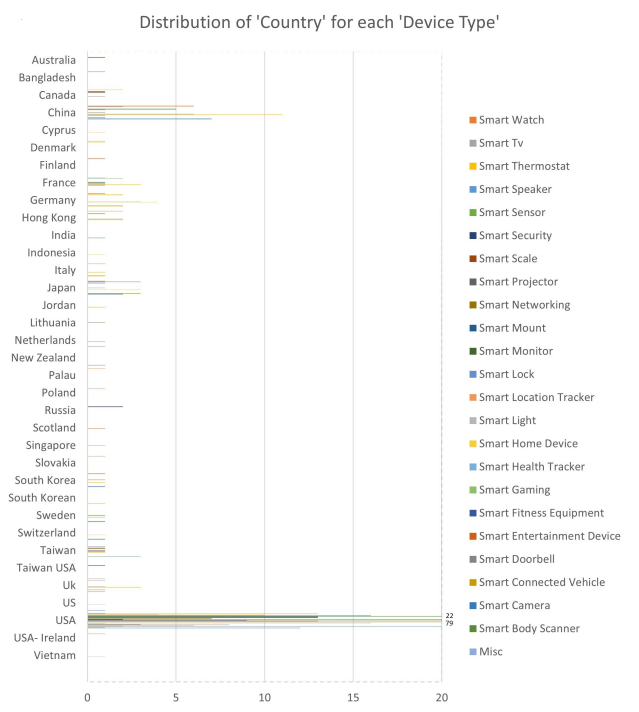


Figure 4: Distribution of Country for each Device Type

privacy policy. We found that the USA was the most prominent manufacturer, with 251 devices, smart home devices constituting the significant chunk (31%) and smartwatches making the smallest chunk (0.40%). China was the second biggest manufacturer with 46 devices, and the most significant chunk comprised smart home devices (23.4%) and the smallest chunk contained smart scales (2.17%). Germany, Japan, France, and the UK also had significant contributions to the list of devices studied.

Furthermore, we discovered that certain categories of devices were more heavily manufactured than others. Smart cameras, smart lights, smart watches, smart sensors, and smart home devices were among the most manufactured devices, with smart home devices being the most prominent. In contrast, smart gaming equipment, smart projectors, smart body scanners, and smart mounts were among the least manufactured categories. We observed that there were fewer options available for users interested in purchasing products in the former categories, indicating that the number of manufacturers could impact the quality of privacy policies.

Our analysis also found that the policies for categories with fewer manufacturers were significantly worse than those with a larger number of manufacturers. We believe this could be due to the former category being more of a monopoly, which may lead to less effort being put into the policies. We found a clear correlation between the number of manufacturers and the quality of privacy policies, which supports our hypothesis.

Our analysis revealed significant gaps in the information provided by privacy policies for smart devices. Specifically,

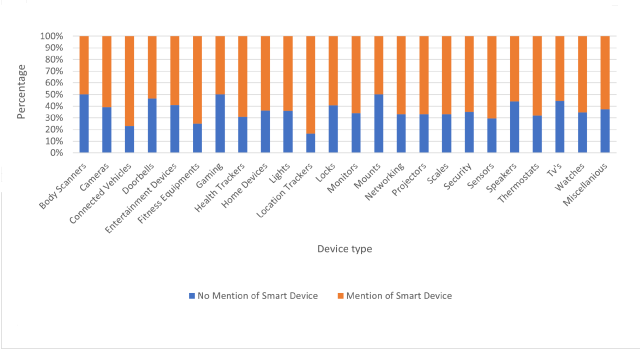


Figure 5: Percentage of Policies with Mention and No Mention of Devices

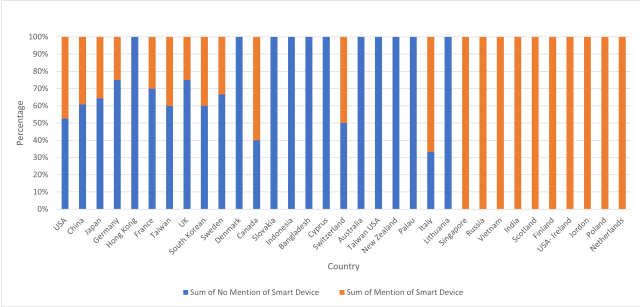


Figure 6: Percentage of Countries with Mention and No Mention of Devices

we found that the absence of mention of devices in privacy policies was particularly high among European manufacturers, with a staggering 60% of countries not mentioning how they handle data collected from devices. This finding indicates that privacy policies need to provide users with more information on how these devices collect data.

Furthermore, while the United States was the largest provider of IoT devices in our study, we found that around 58% of the policies did not mention the devices they handle. Meanwhile, Chinese manufacturers, which are often associated with security concerns [42], actually had a higher rate of mentioning smart devices in their policies, with almost 62% of their policies mentioning smart devices.

It is important to note that just because a privacy policy mentions a device does not necessarily mean that the device is not collecting more information than what is stated in the policy. Therefore, we cannot conclude whether or not the device is "spying" on users based solely on the language used in the policy.

Finally, we categorized the devices into two groups based on the severity of their impact on privacy: critical and less critical devices. Our analysis showed that a majority of critical devices, such as cameras, doorbells, connected vehicles, and location trackers, did not mention their devices and how they collect information. This is particularly concerning as these devices often process vital personal information. On the other hand, less critical devices, such as mounts, projectors, and

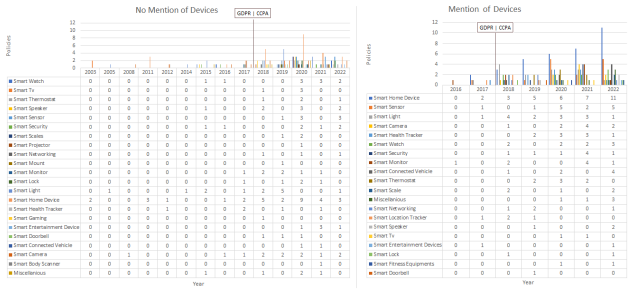


Figure 7: Policies Last Updated

fitness equipment, had higher rates of mentioning their devices in their privacy policies.

4.3.1 Policy Updates

Policy updates play a critical role in ensuring that users are informed of any changes in how their data is collected, processed, and used. In this analysis, we categorized the policies into two groups: device-specific policies and non-device-specific policies. Out of the 462 policies analyzed, only 284 mentioned the last updated policy information. This finding is concerning given that IoT devices store sensitive user information that should be governed by the latest laws and regulations. Failure to comply with these laws and regulations can result in users losing control over their data and a higher risk of data misuse by manufacturers. We then categorized the 284 policies into device-specific and non-device-specific policies, referred to as IoT policies and non-IoT policies, respectively. We found that only one IoT policy was last updated before 2017, while 20 non-IoT policies were last updated between 2003-2016. Additionally, the average number of policy updates from 2017-2022 was higher in IoT policies compared to non-IoT policies. Specifically, 40 IoT policies were updated in 2022, compared to only 18 non-IoT policies. This indicates that IoT policies were updated more frequently, making them more user-friendly and aligned with the latest developments in data privacy laws.

Furthermore, we observed that approximately 88% of the policies were updated after the implementation of GDPR [43] and CCPA [44] in 2018. This could be attributed to manufacturers realizing the need to implement new principles and obligations introduced by GDPR and CCPA, leading to a wave of privacy policy updates. Overall, these findings underscore the importance of policy updates and keeping policies up-to-date with the latest laws and regulations to ensure that user data is protected and handled appropriately.

4.4 Keyword Analysis

In this section, we discuss the keyword analysis conducted as part of our study to gain insights into the privacy policies of various smart devices. In our study, we conducted a key-

word analysis of the privacy policies of various smart devices to gain insights into their privacy practices. By organizing the keywords found in each policy into distinct groups and analyzing their frequency across all the devices studied, we were able to identify trends and patterns that provide valuable insights into the current state of privacy policies in the smart device industry.

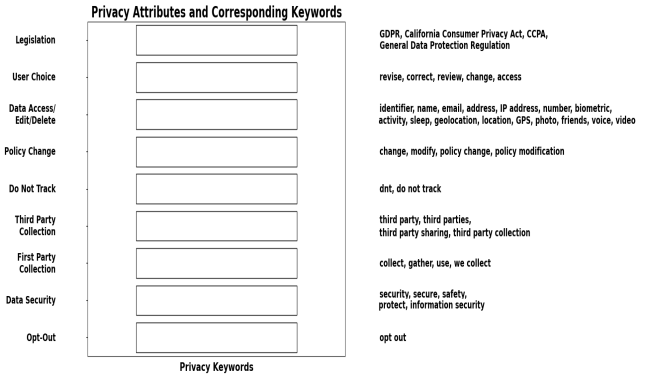


Figure 8: Privacy Attributes and Corresponding Keywords

In order to conduct our analysis, we first organized the keywords found in each privacy policy into ten distinct groups, each group representing specific information about the policy. These groups were identified and listed in Figure 2. Once we had identified the groups, we searched each policy to extract the relevant keywords. Next, we obtained the Maximum, Minimum, and Average values for each group across all the devices we studied, in order to gain a comprehensive understanding of the policies.

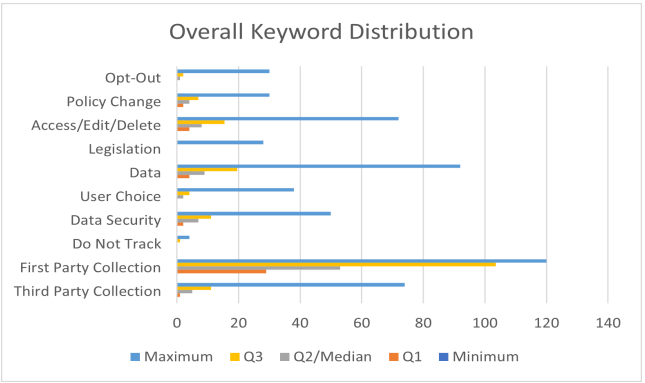


Figure 9: Privacy Attributes and Corresponding Keywords

In the following subsections, we continue to examine each group of keywords individually, discussing their frequency and distribution across the policies of various smart devices and their implications for user privacy and data protection. Through this approach, we provide a comprehensive analysis of the privacy policies of smart devices and the extent to which they prioritize user privacy and data protection.

4.4.1 OPT-OUT

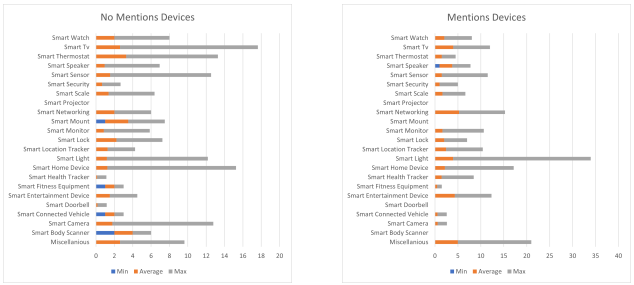


Figure 10: Images illustrating the comparison of Opt-Out procedure values and number of devices by category

The analysis revealed a significant disparity in the frequency of opt-out procedures keywords among the privacy policies of different smart devices. Surprisingly, the lowest mentions of opt-out procedures were observed across all the groups. However, policies that mentioned devices provided a higher average of opt-out procedures compared to those that didn't mention devices. This finding indicates that the inclusion of devices in the privacy policy encourages the provision of opt-out procedures.

Further analysis of the policies that didn't mention devices revealed that smart TVs had the highest recognition for opt-out procedures mentions, while smart doorbells and smart health trackers had the lowest. On the other hand, smart lights had the highest recognition for policies that mentioned devices, while smart fitness equipment had the lowest.

The average number of opt-out procedure mentions across policies that didn't mention devices for minimum, average, and maximum were found to be 0.21, 1.47, and 5.56, respectively. In contrast, policies that mentioned devices had an average of 0.04, 1.92, and 6.65 for minimum, average, and maximum mentions of opt-out procedures, respectively.

In conclusion, the findings suggest that the inclusion of devices in privacy policies may encourage the provision of opt-out procedures. Additionally, the analysis highlights the variability in the mention of opt-out procedures across different smart devices, emphasizing the need for standardization and improved privacy policies.

4.4.2 First Party collection

The present study analyzed the first-party data collection procedures mentioned in various privacy policies, with a focus on the impact of the loss of third-party cookies on device manufacturers. Our analysis revealed that first-party collection had the highest distribution of keywords, indicating that device manufacturers are prioritizing first-party data strategies in the wake of privacy concerns and regulatory laws.

The loss of third-party cookies [45] posed particular challenges for marketers, who initially expressed anxiety about how to track and understand the buyer behaviours of their

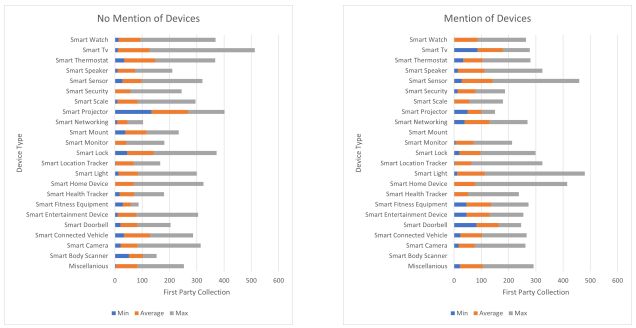


Figure 11: Images illustrating First Party collection across smart device policies

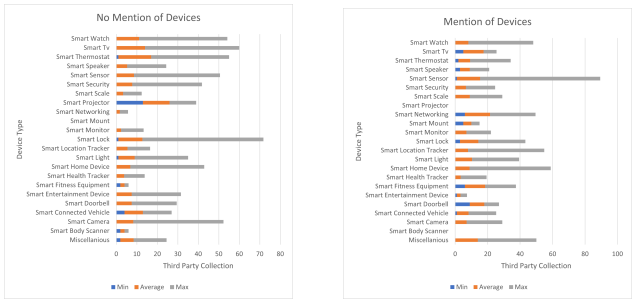


Figure 12: Images illustrating Third Party Collection across smart device policies

target markets. However, marketers have proven to be adaptable and resilient, and they have responded by adjusting their processes and aligning their strategies with new and improved tactics.

Our analysis also revealed that policies that mentioned devices had a higher average for the min, average, and max first-party collection procedures (1.82, 8.05, 22.73) compared to policies that did not mention devices (1.13, 7.04, 22.82), indicating that more policies that mentioned devices provided first-party collection procedures.

Furthermore, among the policies that did not mention devices, policies with smart locks had the highest mentions, while smart fitness equipment and smart body scanners had the lowest mentions. In contrast, smart sensors had the highest mention for policies that mentioned devices, while smart entertainment devices had the lowest mention.

These findings provide valuable insights into the current trends and practices of first-party data collection procedures in the context of privacy concerns and regulatory laws. The study underscores the importance of device manufacturers adopting first-party data strategies to comply with privacy laws and ensure the protection of user data. Furthermore, the study highlights the need for policymakers and device manufacturers to develop effective and transparent data collection policies that prioritize user privacy and data protection.

4.4.3 Third Party Collection

The use of third-party cookies has long been a common practice for collecting user information and providing insights into user behaviour on websites and devices, allowing manufacturers to improve marketing and sales strategies. However, with the enactment of the third-party cookie ban policy in January 2020 [45], the collection of third-party data has been considerably impacted. In our analysis of privacy policies for various smart devices, we found that third-party collection had the sixth lowest mentions overall.

Upon further investigation, we discovered that almost all vendors have blocked third-party cookies in response to the ban, resulting in minimized data collection. In terms of policies that mentioned devices, we found that the average for minimum, average, and maximum third-party collection procedures was lower (23.17, 71.44, 164.17) compared to policies that did not mention devices (23.04, 72.88, 173.04). This indicates that more of the policies that did not mention devices provided third-party collection procedures, possibly because cookies are more prominently used with websites than devices.

Furthermore, among the policies that did not mention devices, smart TVs had the highest mentions for third-party collection procedures, while smart fitness equipment had the lowest mentions. For policies that mentioned devices, smart lights had the highest mentions, while smart projectors had the lowest mentions.

These findings shed light on the current state of third-party data collection procedures in the context of the third-party cookie ban policy. The study underscores the impact of the ban on data collection practices and highlights the need for device manufacturers to adopt alternative data collection strategies to comply with the policy while ensuring the protection of user data. Moreover, the study emphasizes the importance of transparency in data collection policies to prioritize user privacy and data protection.

4.4.4 Access/Edit/Delete

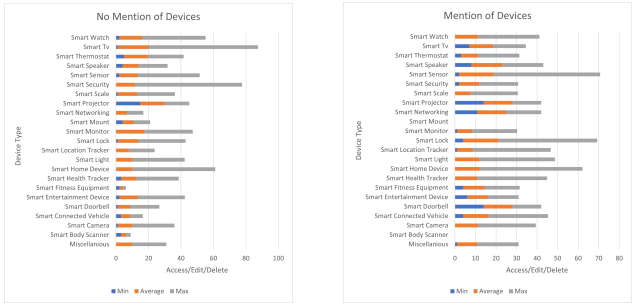


Figure 13: Images illustrating Access/Edit/Delete across smart device policies

This section of our study analyzed the Access/Edit/Delete section of the privacy policies across various smart devices to gain insights into the control measures provided to users regarding their data. Our analysis revealed that the Access/Edit/Delete section had the third highest distribution among the groups, indicating that all the policies in our corpus provided some form of control measures to users on their data.

Further analysis revealed that policies that mentioned devices had a higher average for the min, average, and max Access/Edit/Delete control measures (3.56, 10.57, 24.47) compared to policies that did not mention devices (2.17, 10.34, 26.04), indicating that more policies that mentioned devices provided Access/Edit/Delete control measures.

Furthermore, among the policies that did not mention devices, policies with smart TVs had the highest mentions of Access/Edit/Delete control measures, while smart fitness equipment had the lowest mentions. In contrast, smart sensors had the highest mention for policies that mentioned devices, while smart monitors had the lowest mention.

These findings provide valuable insights into the current trends and practices of Access/Edit/Delete control measures in the smart device industry. The study highlights the importance of device manufacturers providing effective control measures to users over their data to comply with privacy laws and ensure the protection of user data. Additionally, the study emphasizes the need for policymakers and device manufacturers to develop clear and comprehensive data collection policies that prioritize user privacy and data protection.

4.4.5 Data

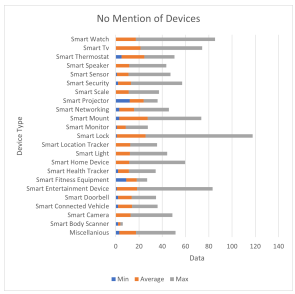


Figure 14: Data

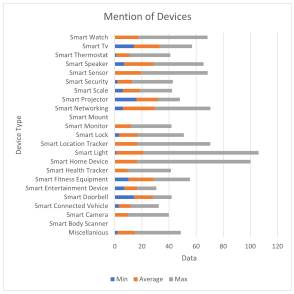


Figure 15: Data

Figure 16: Images illustrating data practices across smart device policies

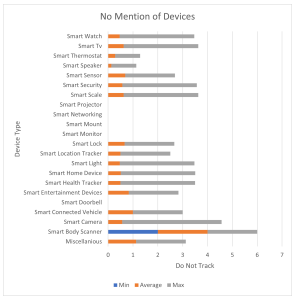
The data section of our analysis sheds light on the information collected by smart device policies, which is crucial for understanding the privacy practices of device manufacturers. Our analysis revealed that the second-highest distribution of keywords belonged to the Data group, indicating that all policies in our corpus collect some form of user information, such as email, ID, location, and name.

Digging deeper into the data, we found that policies that mentioned devices had a higher average for the min, average, and max data collection procedures (4, 13.52, 33.08) compared to policies that did not mention devices (2.13, 13.40, 34.73), highlighting the fact that policies that mentioned devices tend to have more procedures to collect user information.

Furthermore, among the policies that did not mention devices, smart locks had the highest mentions for data collection procedures, while smart body scanners had the lowest. In contrast, smart lights had the highest mentions for policies that mentioned devices, while smart entertainment devices had the lowest.

These findings underscore the importance of understanding the data practices of smart devices and the need for transparent and effective data collection policies that prioritize user privacy and data protection. The insights gained from this analysis can help device manufacturers and policymakers develop policies and practices that ensure user privacy and data protection in the increasingly connected world of smart devices.

4.4.6 Do Not Track



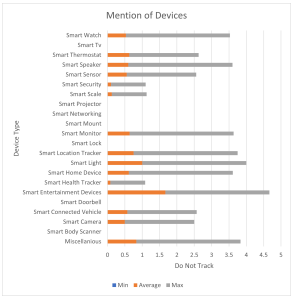


Figure 17: Images illustrating the inclusion of Do Not Track Signal in smart device policies

In the present study, we focused on analyzing the presence of Do Not Track signals in the privacy policies of various smart devices. We found that the lack of Do Not Track signals in privacy policies can be a serious concern as it can affect users in various areas, such as employment, housing, insurance, financial transactions, and government surveillance. Our analysis revealed that the majority of the policies did not provide Do Not Track signals, and the distribution was almost the same as that of Opt-out procedures, making it the second lowest in the group.

Furthermore, we found that policies that mentioned devices had a higher average for the min, average, and max Do Not Track signal procedures (4, 13.52, 33.08) compared to policies that did not mention devices (2.13, 13.40, 34.73), indicating that more policies that mentioned devices had procedures for collecting user information.

Additionally, among the policies that did not mention devices, policies with smart body scanners provided the highest mentions of Do Not Track signals, while smart speakers had the lowest mentions. In contrast, among policies that mentioned devices, smart entertainment devices had the highest mention, while smart security devices had the lowest mention.

The findings of our study highlight the need for device manufacturers to include Do Not Track signals in their privacy policies to ensure the protection of user data and privacy. Moreover, policymakers need to develop more effective and transparent data collection policies that prioritize user privacy and data protection in the smart device industry.

4.4.7 Data Security

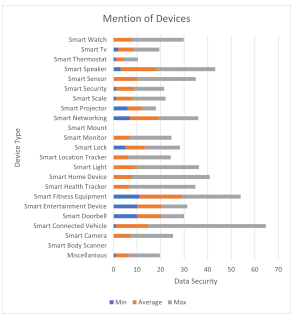


Figure 18: Data Security

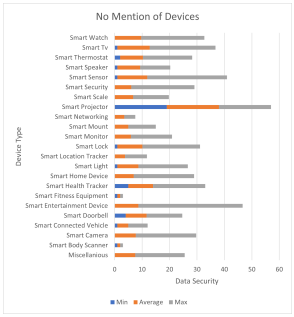


Figure 19: Data Security

Figure 20: Images illustrating Data Security across smart device policies

In today's world, data has become a valuable asset, and it is critical to protect sensitive information from potential breaches. In this study, we conducted a keyword analysis of the privacy policies of various smart devices to gain insights into their data security practices. The Economist [46] calls this era the "zettabyte age," where data centres and the cloud are inundated with an unprecedented volume of information. As a result, stricter privacy laws and regulations have been introduced, and sensitive data must be protected while being stored and transferred.

Surprisingly, our analysis found that data security keywords had the fourth-highest distribution among the groups, even though one would expect it to be a top priority. In our study, we found that policies that mentioned devices had a higher average for the min, average, and max data security procedures (2.56, 8.07, 17.60) compared to policies that did not mention devices (1.65, 7.35, 16.39), indicating that more policies that mentioned devices provided data security measures.

Moreover, among the policies that did not mention devices, policies on smart projectors provided the highest mentions, while smart body scanners had the lowest mentions. In contrast, smart cameras had the highest mention for policies that mentioned devices, while smart thermostats had the lowest mention.

The study's findings underscore the need for device manufacturers and policymakers to prioritize data security measures to protect sensitive information from potential breaches. Device manufacturers must ensure that their privacy policies comply with the latest privacy laws and regulations and provide users with comprehensive information about data security procedures. Policymakers must continue to develop effective and transparent data security policies to safeguard user data and ensure user privacy.

4.4.8 Policy Change

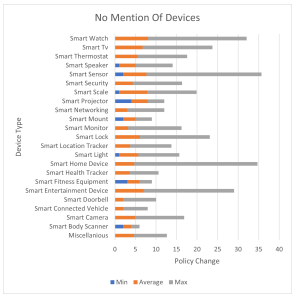


Figure 21: Policy Change

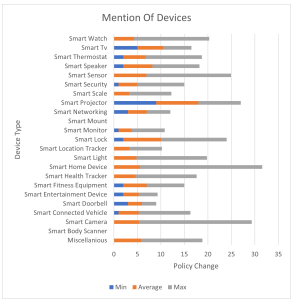


Figure 22: Policy Change

Figure 23: Images illustrating policy change across smart device policies

In this section, we conducted an analysis of privacy policy changes and their disclosure across various smart devices. Our analysis revealed that there is no one-size-fits-all solution for determining when a privacy policy change will occur, as the circumstances can vary widely between different manufacturers and devices. However, the legitimacy and reputation of privacy regulating organizations play a crucial role in ensuring that manufacturers take privacy policies seriously and adhere to the latest regulations.

We found that policies that mentioned devices had a higher average for min, average, and max values (1.43, 4.51, 10.43) compared to policies that did not mention devices (0.695652174, 4.47682608, 12.13), indicating that policies that mentioned devices were more likely to inform users about privacy changes.

Furthermore, among policies that did not mention devices, smart home devices provided the highest mentions of privacy policy changes, while smart cameras had the lowest mentions. For policies that mentioned devices, smart sensors had the highest mention of privacy policy changes, while smart body scanners had the lowest mention.These findings provide valuable insights into the current trends and practices of privacy policy changes and their disclosure in the smart device industry. The study underscores the importance of transparency and accountability in privacy policies and highlights the need for policymakers and manufacturers to work together to develop effective and comprehensive privacy policies that prioritize user privacy and data protection.

4.4.9 Legislation

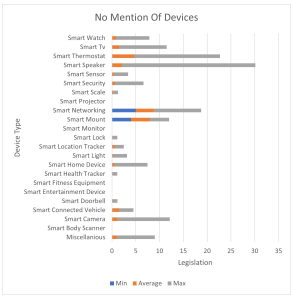


Figure 24: Policy Change

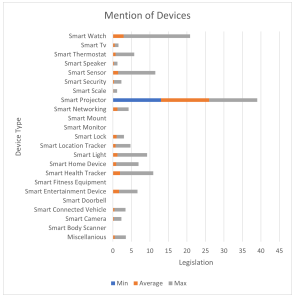


Figure 25: Policy Change

Figure 26: Images illustrating changes in privacy policy across smart devices

Data privacy laws play a crucial role in protecting users' private information, and businesses must adhere to these regulations to ensure user data is handled properly. Our analysis focused on the presence of legislation keywords in privacy policies, with a particular focus on policies from the European Union (EU) due to their stricter privacy regulations, such as GDPR.

We found that the number of legislation keywords collected from EU policies was significantly higher than those from non-EU policies. Additionally, policies that mentioned devices had a higher average for the min, average, and max legislation keywords compared to policies that did not mention devices, indicating a greater emphasis on data protection principles.

Regarding specific devices, policies on smart speakers provided the highest mentions of legislation keywords for policies that did not mention devices, while policies on smart locks had the lowest mentions. For policies that mentioned devices, smart projectors had the highest mentions, while smart speakers had the lowest mentions.

These findings suggest that device manufacturers are placing greater emphasis on data protection principles in their privacy policies, particularly in the context of the EU's stricter privacy regulations. It highlights the importance of policymakers and regulatory bodies in enforcing data protection laws and ensuring user privacy is protected. Device manufacturers must continue to prioritize data protection principles to ensure they are in compliance with data privacy laws and earn the trust of their users.

Overall, it can be observed that policies that mention devices provide a higher keyword count, which helps users navigate the information repeatedly, contributing to better controls for users, privacy regulations, and transparency.

4.5 Checking Similarity of Privacy Policies Before and After GDPR

The General Data Protection Regulation (GDPR) is a comprehensive privacy regulation that enhances individuals' control over their personal data and requires organizations to be transparent about how they collect and use that data [47]. Many organizations have updated their privacy policies to comply with the new regulations, but it is unclear how these policies have changed and whether they are more specific to the organization's data collection and use practices.

To evaluate the impact of GDPR on privacy policies, one approach is to compare the similarity of policies before and after GDPR. This can provide insights into how organizations are responding to the new regulations and how privacy policies are evolving over time. Machine learning models can be used to analyze the similarity of privacy policies and provide an objective measure of their similarity.

One machine learning model that can be used for this purpose is the Sentence-Transformer model, which is trained on a large dataset of text sentence pairs to predict their similarity [48]. The STS-B (Semantic Textual Similarity Benchmark) dataset is a commonly used dataset for training and evaluating such models [49]. The model can be used to compute the similarity score between two text sentences and can be extended to compare the similarity of entire documents, such as privacy policies. To apply this model to compare privacy policies

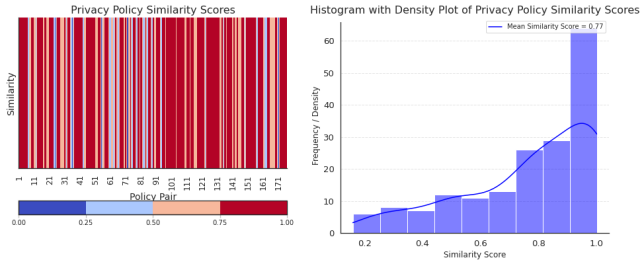


Figure 27: Privacy Policy Similarity Scores. This figure shows a comparison of two visualizations of the similarity scores between privacy policies of IoT devices. The left subplot shows a heatmap that illustrates the similarity scores between different policy pairs for each country, with a color gradient ranging from blue (low similarity) to red (high similarity). The right subplot shows a histogram with a density plot that displays the distribution of the similarity scores across all countries. The mean similarity score is also shown in the legend.

before and after GDPR, we collect the privacy policies of a set of websites before and after GDPR using our framework. The framework utilizes web scraping techniques to collect the policies from the Wayback Machine archives [50]. We also clean the text data to remove irrelevant sections such as headers, footers, and legal disclaimers. We select a representative sample of websites from different industries to ensure the analysis is comprehensive and not limited to a specific sector. The selected websites must have both pre-GDPR and

post-GDPR privacy policies available.

We then compare the similarity scores for each pair of policies using the trained Sentence-Transformer model. If the similarity scores are higher for pairs of policies before GDPR compared to after GDPR, this could indicate that the policies have become more distinct and specific to the organization's data collection and use practices. Conversely, if the similarity scores are lower, it could indicate that organizations are adopting a more uniform approach to privacy policy design to comply with GDPR.

However, it is important to note that the model's performance may be affected by the quality and diversity of the data used to train it, as well as the choice of hyper-parameters during training. Additionally, the model is trained on sentence-level similarity, so it may not be able to capture differences in structure and organization between entire privacy policies. Therefore, it is important to validate the results with a human-in-the-loop evaluation to ensure the model's output aligns with human judgment.

Despite these limitations, using machine learning models to analyze the impact of GDPR on privacy policies can provide valuable insights into how organizations are responding to the new regulations and how privacy policies are evolving over time. This approach can complement traditional legal and policy analyses, and help to identify emerging trends and best practices in privacy policy design and implementation.

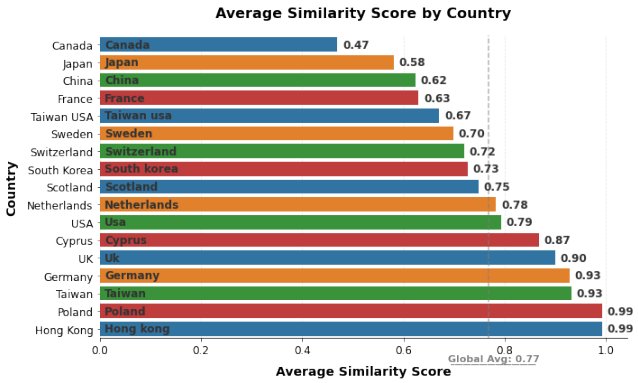
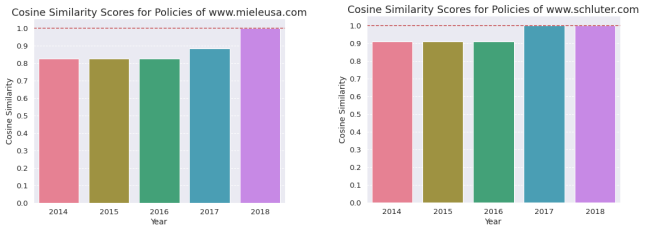


Figure 28: Distribution of Similarity Scores by Country

Our study provides evidence that the impact of GDPR on privacy policies is not uniform across countries. Our analysis of the similarity scores of pre-GDPR and post-GDPR policies found that certain countries, such as the United States, Sweden, and Switzerland, exhibited a higher level of consistency and uniformity in their privacy policy design. In contrast, countries such as Poland and Hong Kong showed a high level of similarity in their policies, indicating that they may have adopted a more uniform approach to privacy policy design to comply with GDPR.

The United Kingdom, on the other hand, had the highest average similarity score, indicating a more nuanced and specific approach to privacy policy design that balances GDPR compliance with organizational data collection and use practices. Similarly, Taiwan had a relatively high similarity score, suggesting a similar trend.



(a) Cosine similarities across for policies of Mieleusa

(b) Cosine similarities across for policies of Schluter

Figure 29: Visualizations of similarity across policies

The global average similarity score for privacy policies before and after GDPR was 0.77. This indicates that, on average, there was a moderate level of similarity between privacy policies before and after GDPR, suggesting that organizations have made some changes to their policies to comply with the new regulations, but there is still room for improvement.

After investigating the impact of GDPR on privacy policies, we wanted to delve a little deeper to understand if there was a specific time period during which policy changes took place for GDPR adoption. Our framework's analysis, as illustrated in Figure 29, revealed that 2017 was a significant year for policy changes, indicating that companies began adapting their privacy policies in response to the GDPR during this time. This finding demonstrates the potential of our framework to provide valuable insights into the evolution of privacy policies, enabling researchers and policymakers to better understand the impact of regulations on organizations and the measures taken to comply with legal requirements.

Our results highlight that GDPR has had a significant impact on privacy policy design, but the extent of this impact varies across different regions. Our approach offers a data-driven method to analyze the impact of GDPR on privacy policies and can be extended to other regulatory contexts to analyze the impact of regulations on organizational practices. However, our analysis also revealed that Canadian organizations have not made substantial changes to their privacy policies following GDPR, indicating a need for further investigation into the reasons for this discrepancy.

In conclusion, our study contributes to the ongoing discussion on the impact of GDPR on privacy policies and emphasizes the importance of considering regional differences in policy design. By leveraging our proposed framework, we can gain a deeper understanding of the factors that drive policy changes and help promote a more transparent and accountable digital landscape, ensuring the protection of user privacy and data.

4.6 Holistic ML Analysis and Mozilla Evaluation

In this section, we concentrate on performing a comprehensive machine learning analysis and assessment using the 172 "Mozilla not found" dataset. We derive this dataset from the Mozilla not found website, which contains information on specific IoT devices and their evaluations. By extracting features such as coherence score, entropy, reading level, and others, we gain valuable insights into the privacy policies of these devices. Our analysis process leverages the data provided by Mozilla and employs our framework to collect the URLs required to access the privacy policies of the corresponding product websites. Our framework aids in collecting and cleaning these privacy policies, allowing us to train a machine-learning model on our original corpus and extract ambiguity measures for the policies. Subsequently, we conduct a keyword analysis on these policies to obtain values related to Opt-Out, Data, Data Security, First-party collection, and others, which we use to create our final dataset.

The Mozilla not found project conducted research on a set of products, examining their privacy policies to evaluate their compliance with the project's standards. The resulting "like" or "dislike" ratings from this assessment served as our target for the final dataset. To achieve a high level of accuracy in our classification model, we opted for logistic regression, which proved to be highly effective when applied to our dataset. As shown in Table 7, our logistic regression-based classification model achieved an impressive accuracy of 93.5%. This demonstrates the suitability of logistic regression in accurately predicting the acceptability of privacy policies based on the Mozilla not found evaluation.

This study demonstrates the efficacy of our framework in collecting, cleaning, and analyzing privacy policies. By combining various analyses, such as keyword analysis, key information analysis and ambiguity extraction, we can generate valuable insights into the quality of privacy policies. The logistic regression model's feature importance analysis in Figure 30 provides insights into the most relevant aspects affecting the classification, with Third Party Collection, Access/Edit/Delete, and User Choice being the top three influential features. Our results also showcase the potential of machine learning models in predicting the acceptability of privacy policies based on the Mozilla not found evaluation.In our analysis, we observed that the prediction performance for class 1 is lower compared to class 0. This can be attributed to the imbalance in the target data, with significantly fewer samples for class 1 (17) compared to class 0 (135). This imbalance may cause the model to be biased towards the majority class, leading to poorer performance in predicting the minority class.

In conclusion, our comprehensive approach to privacy policy analysis, combining machine learning models and various metrics, offers a powerful tool for evaluating and benchmarking privacy policies in the IoT domain. By comparing our results with the Mozilla not found dataset, we further validate our methodology and showcase its applicability in real-world privacy policy evaluation scenarios.

Table 7: Performance Metrics

Class	Precision	Recall	F1-score
Class 0	0.96	0.96	0.96
Class 1	0.67	0.67	0.67

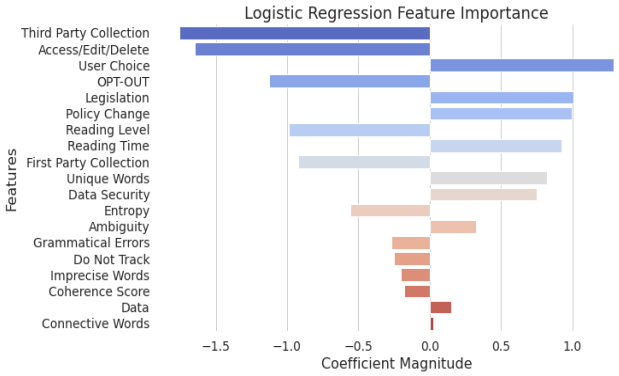


Figure 30: Feature Importance in Logistic Regression Model

5 Conclusion

In this paper, we proposed a comprehensive and automated framework for analyzing and evaluating the privacy quality of IoT device privacy policies. Our framework leverages machine learning algorithms for feature extraction and ambiguity detection and uses web data and historical information to analyze privacy policies across multiple devices.

We applied our proposed framework to a sample of IoT devices and found that our framework can accurately identify key features of privacy policies and detect potential ambiguities that may affect the privacy quality of IoT devices. We also compared our results with those of Mozilla's Privacy Not Included database and found that our framework can provide similar insights into the privacy quality of IoT devices.

While our proposed framework provides a robust approach to analyzing and evaluating the privacy quality of IoT device privacy policies, there are several areas where future research can be focused on to further improve the accuracy and effectiveness of privacy policy analysis. These areas include the integration of natural language processing techniques, the use of more advanced machine learning algorithms, the incorporation of additional data sources, the development of visualization techniques, the extension to other domains, and the exploration of explainable artificial intelligence techniques.

One limitation of our proposed framework is that it relies on publicly available web data and historical information, which may not always be comprehensive or up-to-date. This may

limit the accuracy and effectiveness of privacy policy analysis and require a manual review of certain privacy policies. Additionally, our framework currently focuses on a limited set of privacy features and may not capture all relevant aspects of privacy quality. Future research can explore integrating additional data sources and features to improve the accuracy and comprehensiveness of privacy policy analysis.

Despite these limitations, we believe that our proposed framework provides a valuable contribution to the field of privacy policy analysis and can help to promote greater transparency and accountability in the digital age. By providing users with more nuanced insights into the privacy quality of IoT devices, our framework can help users make more informed decisions about the use of various digital services and protect their personal data and privacy. We hope that our research will inspire further innovation and development in the field of privacy policy analysis and contribute to a more privacy-conscious and responsible digital society.

6 Future Work

While our proposed framework provides a comprehensive and automated approach for analyzing and evaluating the privacy quality of IoT device privacy policies, there are several areas where future research can be focused to further improve the accuracy and effectiveness of privacy policy analysis.

First, future research can explore the use of more advanced machine learning algorithms, such as deep learning or ensemble models, for feature extraction and ambiguity detection in privacy policies. These algorithms can capture more complex relationships between privacy policy features and improve the accuracy of privacy quality assessments. In addition, the integration of natural language processing techniques can help to improve the interpretation and analysis of privacy policies.

Second, future research can investigate the use of additional data sources, such as user reviews or public data breaches, to supplement web data and historical information for privacy policy analysis. Incorporating these sources can provide a more holistic view of privacy quality and identify emerging privacy concerns. Furthermore, the use of explainable artificial intelligence techniques can help to provide users with more transparent and interpretable insights into privacy quality assessments.

Finally, future research can extend our proposed framework to analyze and evaluate the privacy quality of IoT devices in specific domains, such as healthcare or smart homes. This can provide tailored insights for specific use cases and identify domain-specific privacy risks and concerns. Moreover, the framework can be extended to assess the compliance of privacy policies with legal and regulatory requirements, such as GDPR and CCPA.

Overall, future research in these areas can help to further improve the accuracy and effectiveness of privacy policy analysis and address the ongoing challenges associated with privacy protection in the digital age. By incorporating more advanced machine learning algorithms, additional data sources, and interpretable insights into privacy quality assessments, we can provide users with more nuanced and informed insights into the privacy quality of IoT devices. Additionally, extending the framework to specific domains and compliance requirements can help users protect their personal data and privacy while promoting greater transparency and accountability in the digital age.

Availability

All code, framework, diagrams, datasets, and results generated in this study are available on our GitHub page (https://github.com/DAMSlabUMBC/Analysis_Privacy_policies_IoT) for open access and reproducibility. We are also developing a website to provide users with a user-friendly and interactive interface to view the results of our framework. By making our work openly available, we hope to facilitate further research in this area and promote transparency and collaboration in the field of privacy protection in the digital age.

References

[1] Anil Alter, Michele M Tugade, and Barbara L Fredrickson. Smartness as a continuous variable: identifying dimensions of intelligent environments. *Frontiers in psychology*, 7:758, 2016.

[2] Jiaxi Zeng, Xiaoyu Zhang, Dongsheng Luo, and Jianmin Ma. End-to-end user privacy-protecting social recommendation with adversarial training. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1393–1402. ACM, 2017.

[3] Noah Apthorpe, Dillon Reisman, and Nick Feamster. Always on (even when we're off the grid): Privacy risks and conservation benefits associated with the internet of things. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 987–1004. IEEE, 2017.

[4] Paul Biocco, Mahsa Keshavarz, Patrick Hines, and Mohd Anwar. A study of privacy policies across smart home companies. In *An Interactive Workshop on the Human aspects of Smarthome Security and Privacy (WSSP 2018), Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[5] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the international conference on web intelligence*, pages 18–25, 2017.

[6] Barbara Krumay and Jennifer Klar. Readability of privacy policies. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 388–399. Springer, 2020.

[7] Brad Miller, Kaitlyn Buck, and JD Tygar. Systematic analysis and evaluation of web privacy policies and implementations. In *2012 International Conference for Internet Technology and Secured Transactions*, pages 534–540. IEEE, 2012.

[8] Mikhail Kuznetsov, Evgenia Novikova, Igor Kotenko, and Elena Doynikova. Privacy policies of iot devices: Collection and analysis. *Sensors*, 22(5):1838, 2022.

[9] Shomir Liu, Yang Liu, Yuan Li, Shuqin Li, and Fei Niu. Polisis: Automated analysis and presentation of privacy policies using deep learning. *USENIX Security Symposium*, 27(3):531–548, August 2018.

[10] Anantaa Kotal, Anupam Joshi, and Karuna Pande Joshi. The effect of text ambiguity on creating policy knowledge graphs. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1491–1500. IEEE, 2021.

[11] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[12] Neng Wang, Petros Elia, Chun Chen, and Hai Jin. Analysis of online privacy policies using topic models, ambiguity measures, and an entropy model. *Journal of Cybersecurity Education, Research and Practice*, 2019(1):3–22, 2019.

[13] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, 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)*, pages 1330–1340, 2016.

[14] Herman T Tavani. Informational privacy: Concepts, theories, and controversies. *The handbook of information and computer ethics*, pages 131–164, 2008.

[15] Hendrik JG Oberholzer and Martin S Olivier. Privacy contracts as an extension of privacy policies. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1192–1192. IEEE, 2005.

[16] Carlos Jensen, Chandan Sarkar, Christian Jensen, and Colin Potts. Tracking website data-collection and privacy practices with the iwatch web crawler. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 29–40, 2007.

[17] Mukund Srinath, Soundarya Nurani Sundareswara, C Lee Giles, and Shomir Wilson. Privaseer: A privacy policy search engine. In *International Conference on Web Engineering*, pages 286–301. Springer, 2021.

[18] datamation. About Top Iot devices.

[19] mozilla.org. About Smart Home.

[20] *SACMAT '08: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies*, New York, NY, USA, 2008. Association for Computing Machinery.

[21] Apurwa Yadav, Aarshil Patel, and Manan Shah. A comprehensive review on resolving ambiguities in natural language processing. *AI Open*, 2:85–92, 2021.

[22] Alexander Franz. Automatic ambiguity resolution in natural language processing: An empirical approach. 1996.

[23] Roger C. Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631, 1972.

[24] Gonzalo Génova, José Miguel Fuentes, Juan Llorens Morillo, Omar Hurtado, and Valentín Moreno. A framework to measure and improve the quality of textual requirements. *Requirements Engineering*, 18:25–41, 2011.

[25] Anantaa Kotal, Karuna Pande Joshi, and Anupam Joshi. Vicloud: Measuring vagueness in cloud service privacy policies and terms of services. In *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*, pages 71–79, 2020.

[26] Joel R. Reidenberg, Jaspreet Bhatia, Travis D. Breaux, and Thomas B. Norton. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163 – S190, 2016.

[27] Julien B Kouame. Using readability tests to improve the accuracy of evaluation documents intended for low-literate participants. *Journal of MultiDisciplinary Evaluation*, 6(14):132–139, 2010.

[28] Rudolf Flesch. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007, 2007.

[29] Daniel Naber et al. A rule-based style and grammar checker. 2003.

[30] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[31] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[32] Elfrieda H Hiebert. Unique words require unique instruction.

[33] James Hiebert and Anne K Morris. Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of teacher Education*, 63(2):92–102, 2012.

[34] Adam Geitgey. Natural language processing is fun! *Medium, July*, 18, 2018.

[35] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[36] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. IEEE, 2017.

[37] Hua Yu and Jie Yang. A direct lda algorithm for high-dimensional data—with application to face recognition. *Pattern recognition*, 34(10):2067–2070, 2001.

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[39] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[40] David M.W. Powers. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Springer, 2011.

[41] Didier Masha, Lek Chaisorn, and Santi Phithakkitnukoon. Smart homes privacy policies analysis: A critical information analysis. *Journal of Ambient Intelligence and Humanized Computing*, 11(10):3975–3987, 2020.

[42] James Griffiths. Us warns that chinese-made drones could be giving spy agencies intel. *CNN*, 24, 2019.

[43] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[44] California consumer privacy act, 2021.

[45] Cookiebot. Google third-party cookies: How they work and how to opt-out. https://www.cookiebot.com/en/google-third-party-cookies/, n.d. Accessed March 10, 2023.

[46] The Economist. The data deluge. *The Economist*, 411(8884):12–15, 2014.

[47] Peter Voigt and Arndt von dem Bussche. The eu general data protection regulation (gdpr): A practical guide. *Springer*, 2017.

[48] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[49] Daniel Cer, Mona Diab, Eneko Agirre, Iñaki Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, 2017.

[50] Brewster Kahle and Rick Weber. The Wayback machine: The web archive for preservation and research. In *Proceedings of the 9th Web Archiving and Digital Libraries Conference*, Baltimore, Maryland, USA, 2010.