Computers, Materials & Continua
DOI: 10.32604/cmc.2023.034039
Article

A GDPR Compliant Approach to Assign Risk Levels to Privacy Policies

Abdullah R. Alshamsan[1] and Shafique A. Chaudhry[1,2,*]

[1]Department of Computer Science, Clarkson University, Clarkson Ave, Potsdam, 13699, NY, USA
[2]David D. Reh School of Business, Clarkson University, Clarkson Ave, Potsdam, 13699, NY, USA
*Corresponding Author: Shafique A. Chaudhry. Email: schaudhr@clarkson.edu
Received: 05 July 2022; Accepted: 20 September 2022

Abstract: Data privacy laws require service providers to inform their customers on how user data is gathered, used, protected, and shared. The General Data Protection Regulation (GDPR) is a legal framework that provides guidelines for collecting and processing personal information from individuals. Service providers use privacy policies to outline the ways an organization captures, retains, analyzes, and shares customers' data with other parties. These policies are complex and written using legal jargon; therefore, users rarely read them before accepting them. There exist a number of approaches to automating the task of summarizing privacy policies and assigning risk levels. Most of the existing approaches are not GDPR compliant and use manual annotation/labeling of the privacy text to assign risk level, which is time-consuming and costly. We present a framework that helps users see not only data practice policy compliance with GDPR but also the risk levels to privacy associated with accepting that policy. The main contribution of our approach is eliminating the overhead cost of manual annotation by using the most frequent words in each category to create word-bags, which are used with Regular Expressions and Pointwise Mutual Information scores to assign risk levels that comply with the GDPR guidelines for data protection. We have also developed a web-based application to graphically display risk level reports for any given online privacy policy. Results show that our approach is not only consistent with GDPR but performs better than existing approaches by successfully assigning risk levels with 95.1% accuracy after assigning data practice categories with an accuracy rate of 79%.

Keywords: GDPR; machine learning; natural language processing; privacy assessment; privacy policy; text classification

1 Introduction

Internet users access various websites and applications in to avail themselves of services from various enterprises, such as banking, healthcare, entertainment, etc. Most of these websites collect information from users. Organizations are required to ensure that user data is collected, stored, and shared in an ethical and legal manner. International law and regulatory bodies such as the

Federal Trade Commission (FTC) [1], the Organization for Economic Co-operation and Development (OECD) [2], and the General Data Protection Regulation (GDPR) [3] require service providers to inform users about how their data is protected. In addition to the international regulations, various other laws have been passed in different regions and countries, e.g., [4,5]. As evidence that they are compliant with such regulations, the majority of service providers present the terms of their services, data protection and usage practices, and other contractual agreements in privacy policies.

These policies can be quite intrusive, yet inform the users about the conditions under which their data is gathered, retained, used, and shared with partners or sold to other companies. Most users agree to such policies without realizing the risks to their data [6]. Studies have shown that most users do not read privacy policies [7,8]. Firstly, the policies use a lot of legal jargon, making them difficult for a layperson to understand. Secondly, they are complex and lengthy [9]. It has been shown that the average time needed to read one privacy policy is around 18 min [10]. On average, for a user to read the policies for all of the services they use would take about 201 h per year [8]. One approach to help customers is to automate the process of creating a summary of privacy risks associated with the practices described in a policy.

Several research efforts have been proposed in the literature to automate policy understanding. These vary from providing only a text summary of the policy [11,12] to assigning risk levels to various practices used by the service provider. Recently, Machine Learning (ML) techniques [11,13–15] have attracted a lot of interest from the research community to identify privacy practices from policies and/or assign risk levels to various sections of policies. Most of these ML techniques use training datasets, which consist of privacy policies that are manually annotated/labeled to classify the data practices into certain categories. Once the model is trained, it is used to make predictions for the test dataset. The labeling (annotation) activity to assign risk levels to data practices in policies is time-consuming and costly, however; and no research so far has focused on the issues of cost, time, and the mistakes made during manual annotation. Moreover, there is a lack of standard presentation in the existing approaches, meaning that each project has involved privacy experts in writing the privacy aspects, risk factors, and description of data protection regulations adopted in the research approach, rather than a standard set of regulations.

We present a framework that helps users see not only the compliance of data practices with GDPR but also the privacy risk levels associated with accepting that policy. Our approach uses an ML model to classify each sentence/segment of the privacy policy content according to one of ten data practice categories described in [16], in common with existing approaches, but our risk assignment technique is novel. The main contribution of our approach is mitigating the overhead cost of human efforts for dataset annotation by using the most frequent words in each category to create word-bags, which are then used with Regular Expressions (RegEx) and Pointwise Mutual Information (PMI) scores to assign risk levels that comply with the GDPR guidelines for data protection. The use of RegEx quadgrams and PMI scores assigns risk levels automatically without the human effort required for manual annotation. As a result, our approach does not require the training time that is needed for ML techniques that manually annotate a dataset. To the best of our knowledge, our work is the first to use this method in this domain. Results show that our approach is not only consistent with GDPR, but it also performs better than existing approaches by successfully assigning the risk levels with 95.1% accuracy after assigning the data practice categories with an accuracy rate of 79%.

The rest of the paper is organized as follows: Section 2 provides an overview of GDPR. A review of the state-of-the-art in this domain is presented in Section 3 followed by our approach in Section 4.

In Section 5, we demonstrate our visual report of the risk level and present the evaluation results in Section 6. Section 7 concludes the paper with directions for future work.

2 GDPR and Data Practices

The General Data Protection Regulation is the most recent law regarding data protection and privacy that applies to the European Union. It was adopted on 14 April 2016 and implemented on 25 May 2018. One of its goals is to make users aware of how their data is collected, used, shared, secured, and processed by companies or service providers. Another key goal is to make privacy policies clear, comprehensive, and easy for ordinary users to read and understand. Thus, the GDPR gives people more control over their personal data. The fundamental rules in the GDPR concern the protection of personal data. In particular, Article 12 is the primary source of guidance regarding privacy policies, requiring data processing and communications to use clear and plain language and be transparent, concise, intelligible, and easily accessible [17]. The main GDPR article covering personal data privacy is Article 13, which requires users to be told who is collecting and processing the data, what types of personal data are being collected, for what purposes, and for how long the service provider will store it. In addition, Article 13 states users' rights, such as to request from the service provider access to the data, its erasure, and restrictions on its processing. The user also has the right to complain to a supervisory authority and to transmit and receive their data from another service provider [18].

Liu et al. [19] proposed an approach to check how much a policy's data practices comply with Article 13 of the GDPR. While this is a significant effort, it covers only that article. In another study, Poplavska et al. [20] mapped GDPR articles according to the OPP-115 Corpus (Online Privacy Policies, set of 115) annotation scheme. The correspondences between OPP-115 and the GDPR shown in this work provides a bridge to future research but lacks any mechanism to assign risk levels to policy statements. Table 1 outlines how data practice categories are mapped to GDPR.

Table 1: GDPR articles mapped to OPP-115 categories

OPP-115 category	Description	GDPR articles [20]
First Party Collection/Use	How and why does the (company/website/service provider) collect personally identifiable information (PII)?	4, 5, 6, 7, 8, 9, 10, 11, 24, 25, 30, 33, 34, 35, 36, 37, 38, 39, 89, 91, 95
Third Party Sharing/Collection	How is PII shared with or collected by third parties?	4, 6, 9, 19, 28, 29, 30, 37, 38, 39, 44, 45, 46, 47, 48, 49, 96
User Choice/Control	What control options and choices are available to users?	4, 6, 7, 8, 9, 13, 14, 17, 18, 20, 21, 26, 49, 77, 78, 79, 80, 82
User Access, Editing and Deletion	Can users access, edit, or delete their information? If so, how?	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 25
Data Retention	How long is PII stored?	5, 13, 14, 25, 30
Data Security	How is PII protected?	4, 5, 6, 12, 24, 25, 28, 30, 32, 33, 34, 35, 36, 45, 89

(Continued)

Table 1: Continued

OPP-115 category	Description	GDPR articles [20]
Policy Change	Will users be informed about changes to the privacy policy? If so, how?	12
Do Not Track	Are online tracking and advertising signals honored? If so, how?	
International and Specific Audiences	Practices pertaining only to a specific group of users (e.g., children).	8
Other	Additional practices not covered above, such as contact information.	

3 Related Work

A number of research efforts have explored the problem of assessing the risks associated with data collection, retention, and sharing practices provided in a privacy policy. The main goal of these approaches is to help users understand a service's data practices before signing a contract with the service provider of that service. In this section, we provide a brief overview of several studies that use Natural Language Processing (NLP) and ML techniques to address risk levels in privacy policies.

Wilson et al. [16] created the OPP-115 corpus to mitigate the shortage of datasets that identify data practices in privacy policies. They analyzed policies from 115 websites and evaluated the results using three ML approaches to classify them into ten categories covering most aspects of data collection, retention, and sharing practices. To evaluate the performance of ML models, they used Support Vector Machine (SVM), logistic regression, and hidden Markov models as classification models.

Polisis [13] is a tool that combines a privacy-centric language model and neural networks to analyze privacy policies. It uses the OPP-115 dataset that is annotated by three legal experts to label privacy practices. The Polisis model then assigns privacy icons. However, it uses only five icons, based on the Disconnect icons [21].

A crowd-sourcing approach that uses a shortlist of positive and negative aspects of the privacy policies is presented in [22]. The approach uses online volunteers to read, discuss, and then rate privacy policies in the form of grades A-E, where A is best and E is worst. The overall rating is calculated from all responses provided. This approach has suffered from a lack of participation. Only a handful of website policies have been rated since 2012 and only eleven have been finally judged and given an overall grade. Since companies frequently update their privacy policies and there is no mechanism to keep track of these changes, most of the ratings are outdated.

Privee [23] is a tool that uses the result from [22] if available, with ML and NLP, to categorize policies. If no result is available, ML is used to produce one. This tool achieves an overall F-1 score of 90% and has six categories, the criteria for which are not based on any regulations such as GDPR or OECD. As a result, the annotation criteria and ML are not detailed in this work.

PrivacyGuide [24] is a tool used to summarize and classify privacy policies into 11 categories, then provide an associated risk level for each category. The risk levels are categorized as green, yellow, and

red and the 11 categories are defined according to pre-GDPR criteria and groupings that were used in some previous studies. The privacy experts then devised a rule to assign the risk level for each category (each having three rules, one for each risk level). Each sentence/segment was then manually labeled to the associated risk level. The tool used 45 policies to create the corpus. However, the annotation criteria are not detailed and some significant sentences were omitted, which could make a difference to the general meaning.

PrivacyCheck v2 [25] automatically summarizes any privacy policy by answering 20 questions. Ten questions are about user control and are based on the work of the OECD and the FTC Fair Information Practices (FIP), while the other ten are based on the GDPR. To train this model, the researchers used a corpus from an earlier version of PrivacyCheck [14]. The model contains 400 privacy policies and risk levels (Green, Yellow, and Red) manually assigned by privacy experts. The average accuracy of the GDPR and User Control questions in finding the correct risk level was 60%. PrivacyCheck v2 covers only ten privacy factors related to personally identifiable information, which is quite limited. Moreover, the dataset is less reliable, since only 60 of the 400 privacy policies (15%) were selected to perform the quality control test. The researchers defined new criteria for privacy policies, then created a corpus, but some of these criteria cannot be applied to every privacy policy; for example, the criterion Credit Card Number cannot be applied to a non-finance-related service provider.

Ours is a novel approach using a particular RegEx type called Flexgrams, allowing us to create a quadgram collect set with a PMI score to solve this problem. By merging these techniques, our approach can assign the risk level automatically without the human effort required for manual annotation. Another benefit of our approach is that it solves the problem of frequent updates to the content of privacy policies, since there is no need for human intervention.

Our approach differs from the prior studies in several ways. First, it balances transparency and user consent by using the GDPR to build rules, because it is essential to associate rules with privacy data regulations. In addition, it reduces cost and time because it dispenses with the need for human effort for the manual annotation of risk levels to policy statements. While we use SVM for the categorization of the policy text into different categories, the automated risk assigning component does not use ML and therefore needs no training. The word-bags can be updated easily, which supports flexibility and scalability over time. Moreover, we cover a near-comprehensive range of data practices mentioned in privacy policies. Table 2 compares the features of existing approaches with ours. To the best of our knowledge, no approach to determine the risk levels for privacy policies has been performed using RegEx and PMI scores under the umbrella of the GDPR.

Table 2: Comparison of existing approaches' features with our approach

Features	P1	P2	P3	P4	P5	P6	OAP
Uses GDPR	×	×	×	×	×	✓	✓
Uses a reliable dataset (reviewed by privacy experts)	✓	✓	×	×	✓	×	✓
Number of privacy factors	5	10	6	10 PII		11	10
Assigns risk level automatically (NOT rated by human experts)	×	×	×	×	×	×	✓
Assigns risk levels in real-time (No training needed)	×	×	×	×	×	×	✓

(Continued)

Table 2: Continued

Features	P1	P2	P3	P4	P5	P6	OAP
Accuracy in assigning data practices	81%	66%	N/A	60%	N/A	74%	79%
Accuracy in assigning risk levels	88.4%	N/A	N/A	N/A	N/A	90%	95.1%

LEGEND P1: Polisis, P2: OPP-115, P3: Privee, P4: PrivacyCheck-v2, P5: ToS;DR, P6: PrivacyGuide, OAP: Our Approach, PII: Personally identifiable information.

4 Proposed Approach

In this section, we describe the architecture and components of our approach. The system architecture, shown in Fig. 1, consists of three major modules: the machine learning model to predict the high-level class of privacy policy paragraphs, a word-bag and collection set generator, and a risk level generator that assigns risk levels (low, medium, and high) according to extracted information on the collection set weight.

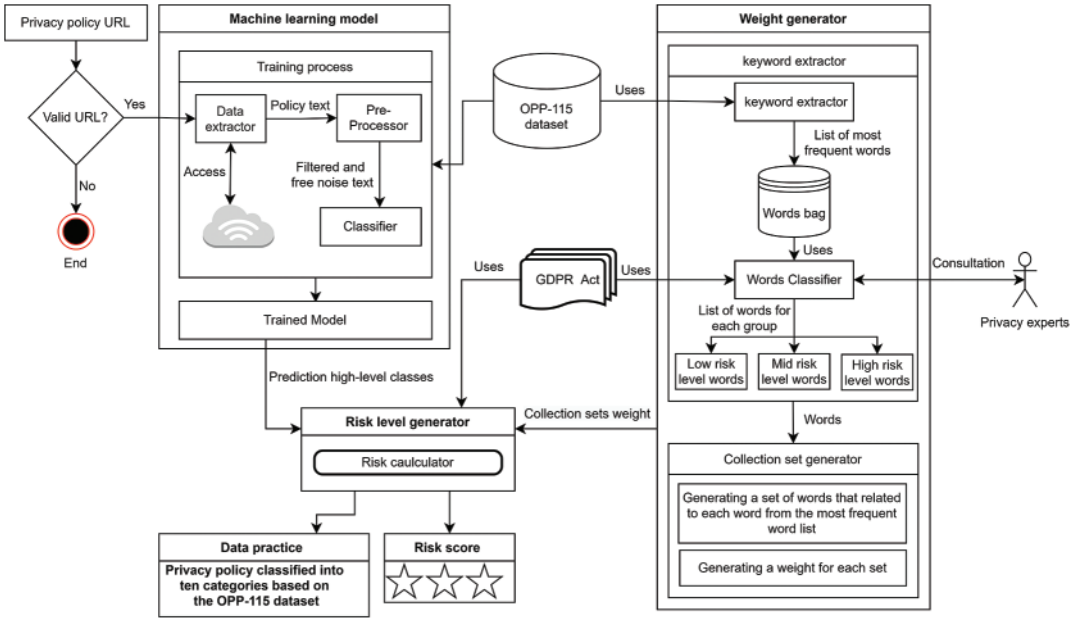

Figure 1: System architecture

4.1 Machine Learning Model

The ML model reads an online policy, provides a high-level prediction for every paragraph, and places the paragraphs into ten high-level classes described in the OPP-115 dataset. This component has three modules: Data extractor, Pre-Processor, and Classifier. The data extractor is used to extract plain text from the website's privacy policy page. The extracted text is used for pre-processing and parsing. The pre-processor module is used to remove special characters and stop words from the policy text as well as to perform tokenization and lemmatization procedures. Tokenization involves

splitting sentences and paragraphs into smaller units called tokens while lemmatization is the process of reducing a given word to its root word, called a lemma. After completion of the pre-processing step, the noise-free text is passed to the ML classifier.

The main goal of the classifier is to place each sentence in the most suitable category. We split the dataset data into 75% for training and 25% for testing purposes and applied various ML models. We compared the performance of ten ML classifiers: Bagging, Decision Tree, Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Multinomial Naive Bayes, Random Forest, Stochastic Gradient Descent, SVM, and XGBoost. The performance was compared on the basis of Precision, Recall, F1-score, and Accuracy scores. These metrics are based on a confusion matrix that comprises true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Accuracy refers to the fraction of correct predictions overall predictions and is given in Eq. (1).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{1}$$

Precision refers to the proportion of correctly predicted positives to all positives, as shown by Eq. (2).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2}$$

Sensitivity, also known as recall, refers to the fraction of known positives that are correctly predicted and is given in Eq. (3).

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

F1-score is the harmonic mean of precision and recall and is given by Eq. (4).

$$\text{F1} - \text{score} = \frac{\text{Precision} * \text{ Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

The SVM model performed the best for accuracy, precision, recall, and F1-score and was selected as our main classifier. Table 3 shows a performance comparison of all the classifiers we tested. Table 4 shows the performance of the SVM classifiers for each of the ten categories. The high-level predicted categories are passed to the risk level generator, where the risk level is assigned.

Table 3: A comparison of all classifiers used

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.76	0.77	0.79	0.77
Support Vector Machine	0.79	0.84	0.77	0.79
Stochastic Gradient Descent	0.75	0.78	0.73	0.75
Bagging	0.73	0.78	0.62	0.67

(Continued)

Table 3: Continued

Classifier	Accuracy	Precision	Recall	F1-score
Multinomial Naive Bayes	0.78	0.79	0.75	0.78
Gradient Boosting	0.73	0.75	0.69	0.71
K-Nearest Neighbors	0.75	0.78	0.68	0.70
Random Forest	0.72	0.81	0.58	0.64
XGBoost	0.71	0.72	0.66	0.68
Decision Tree	0.64	0.6	0.52	0.54

Table 4: SVM model performance

Category	Precision	Recall	F1-score
First Party Collection/Use	0.78	0.83	0.80
Third Party Sharing/Collection	0.81	0.77	0.79
User Choice/Control	0.73	0.57	0.64
User Access, Edit and Deletion	0.68	0.70	0.69
Data Retention	1.00	0.29	0.44
Data Security	0.89	0.82	0.85
Policy Change	0.94	0.91	0.92
Do Not Track	1.00	1.00	1.00
International and Specific Audiences	0.86	0.89	0.87
Other	0.74	0.79	0.76
Accuracy			0.79
Macro avg	0.84	0.76	0.78
Weighted avg	0.79	0.79	0.78

In the work reported in this paper, we did not use Deep Learning (DL) because the performance of DL models was very similar to the traditional ML models in classifying the privacy policy using the OPP-115 dataset. Moreover, traditional ML models take less time to train and have less computational complexity than DL models. Our decision was based on our previous results, which contain several comparison tables [26]. We also selected SVM as our primary classifier, since it has the best performance among all previous studies that used the OPP-115 dataset and traditional ML [16,26–28].

4.2 Weight Generator

The second component in our approach is the weight generator, which can be divided into two sub-components: 1) the keyword extractor and 2) the collection set generator. The keyword extractor extracts the most frequent words in the OPP-115 corpus. Table 5 shows examples of these words. We used the services of multiple privacy experts to classify them into different risk-level groups depending on how each word is associated with a specific risk level. In addition to their knowledge of the GDPR and their expertise in the privacy policies domain, the experts must also fully understand the tasks. Therefore, before starting the main task, we provided the experts with several examples of how a certain article of the GDPR is mapped to the ten data practice categories from the OPP-115 corpus.

Table 5: Examples of most frequent words

Category	Words
First-Party Collection/Use	collect, personal, location, IP address, identifiable, credit card, contact, age, address, account
Third-Party Sharing/Collection	share, sell, disclose, provider, third party, transfer, service providers, marketing partners, subsidiaries, disclosure
User Choice/Control	unsubscribe, disable, choose, choice, consent, option, wish, agree, opt-in, opt-out
User Access, Edit and Deletion	delete, profile, correct, change, update, access, request, modify, preferences, accurate
Data Retention	retain, store, delete, database, participate, record, remove, keep, backup, discard
Data Security	secure, security, safeguard, protect, compromise, encrypt, unauthorized, SSL, encryption, restrict
Policy Change	change, change privacy, policy time, current, policy agreement, update privacy, update, notice
Do Not Track	signal, track, track request, browser, disable, track setting, cookies, web beacons, IP address
Purpose	ads, services, verifying, fraud, prevention, improve products, identification, promotions, advertising, analytics

Table 6 shows an example of mapping the OPP-115 corpus onto GDPR articles. The initial task required the experts to classify the produced frequent words into three groups, each representing one particular risk level. The classification rule finds the probability of a particular word indicating a specific level of risk. Using the GDPR descriptions in the different articles, we calculated the probability for each word and selected several words to classify into the three groups, omitting ambiguous words such as ‘occasional’ and ‘perhaps’. We also suggested adding to each group some necessary words in each group which, while not frequently used, are nonetheless essential, such as ‘opportunity’, ‘decide’, ‘permission’, etc. This task generated three bags of seed keywords for each level of risk. Initially, we selected 142 words from the OPP-115 corpus. After the review and input from the privacy experts, we reduced the list to 118 words divided into three levels of risk: low, medium, and high, comprising 41, 47, and 30 words respectively. The second sub-component is the collection set

generator, which generates a collection set for each word. A collection set is an organized group of objects accumulated in one location for a particular purpose [29].

Table 6: Mapping certain GDPR articles to OPP-155

Part of policy content from OPP-115	"We will retain your information for as long as your account is active or as needed to provide you with services. We will retain and use your information as necessary to comply with our legal obligations, resolve disputes, and enforce our agreements."
Privacy category	Data Retention
GDPR article numbers	5, 25, 30, 13, 14
Text of GDPR articles	Art. 5(E): "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject" Art. 25(2): "The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. 2 That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. 3 In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons." Art. 30(1-f): "where possible, the envisaged time limits for erasure of the different categories of data;" Art. 13/14(2-a): "the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period"
Mapping the GDPR to a Privacy Policy category	Depending on the purposes of holding the personal data from the service provider, the storage period needs to be determined.

Since we were working with multi-class data containing several ambiguous words that could fall into multiple categories, there was a very high chance of similarity between the sets. Therefore, we used n-gram, which is a sequence of n consecutive words. To add a layer of complexity, we chose the n-gram

of n = 4, i.e., sets of four words, which we called quadgrams. These words were presented in the order that they appear in the text because word order is significant for the next phase of this approach. We used quadgrams to ensure that our system covered most aspects of English sentence structure. English has four types of sentence structure: simple, compound, complex, and compound-complex [30]. At the same time as creating the collection set, we calculated the pointwise mutual information (PMI) [31] measure for each collection set. The PMI is a measure of association between two or more words, (a, b), based on the ratio between the co-occurrence probability P(a, b) and the independent probability of observing (a, b) by chance, P(a)P(b) [32]. Eq. (5) was used to calculate the PMI for two words (a, b). Next, we applied Eq. (6) to calculate the PMI. The result was a list of collection sets with their PMI scores for each keyword.

$$\mathrm{PMI}\,(a, b) = \log\left(\frac{p\,(a, b)}{p\,(a)\,p\,(b)}\right) \tag{5}$$

$$\mathrm{PMI}\,(w1, w2, w3, w4) = \log\left(\frac{p\,(w1, w2, w3, w4)}{p\,(w1)\,p\,(w2)\,p\,(w3)\,p\,(w4)}\right) \tag{6}$$

4.3 Risk Level Generator

The third component in our approach is the risk level generator, which is the main and most complicated component, aimed at generating the risk level for a given sentence. We used RegEx as the primary technique to achieve this goal and calculate the risk accurately. The first step in generating the risk levels is finding the intersection between any words in a given sentence and any keyword from the risk level groups. If there is an intersection, we move down to a low level (collection sets). We retrieve all collection sets and PMI scores associated with these words. The next step in this stage is to apply the regular expression rules to check whether all the collection set words are presented in the sentence correctly. If there are more than two sets that have met this condition, our approach will pick the highest PMI score. Also, if the sentence contains words that intersect with different risk level word groups, we repeat all the previous steps for each word and select the highest PMI score. After we have finished all of these steps and have the highest PMI score, we can determine the risk level for this sentence.

To clarify the process, we shall take as an example the following sentence from the text of one of the privacy policies: "We share your information with third parties." This sentence contains seven words. The word 'share' is in the high-risk level group and intersects with other words in the sentence. Therefore, we move down to a low level (collection sets). We retrieve all collection sets and PMI scores associated with 'share', such as ('we', 'share', 'your', 'information') or ('share', 'your', 'information', 'with') etc. If there are a number of sets that can apply to this sentence, we select only the collection set with the highest PMI score and determine the risk based on this score.

To determine the pattern, we use a regular expression rule called Flexgrams, a sequence with one or more variable-length gaps. We also apply position tracking for each word in the sentence to have the power to make some rule exemptions. Algorithm 1 shows the sequence of all these steps and Table 7 shows an example for one sentence. Once the privacy category is predicted using the ML component, the risk level generator component generates the risk level we provide to the end-user with the data practice category.

Table 7: Example for one sentence

Example:	You should receive marketing emails only from us and, if you agree, from other organizations we have carefully chosen		
Keyword	Risk level	Candidate collection set	PMI score
receive	HIGH	('receive', 'from', 'other', 'organization')	22.7246656223
agree	LOW	('if', 'you', 'agree', 'at')	18.8946457672
choose	LOW	('if', 'you', 'have', 'choose')	16.8667397706
Final risk level	HIGH		

Algorithm 1 Risk levels measurement algorithm

Input: A sentence from privacy policy text
Output: Return the risk level of this sentence

```
PMIscore = 0 ; RiskLevel=low ; seeds = The list of risk levels words
for word in all sentence words do
    for keyword in the seeds list do
        if word = keyword then
            for CollectionSetWords in a keyword database do
                if CollectionSetWords ∈ sentence then
                    if CollectionSetWordsPMI > PMIscore then
                        PMIscore = CollectionSetWordsPMI
                        RiskLevel = keyword risk level tag
                    end if
                end if
            end for
        end if
    end for
end for
return RiskLevel
```

5 Visual Report of the Risk Level

Visualizing the privacy policy texts alongside associated risk levels for particular segments helps users focus on the high-risk part only. We developed a report generator to interpret the risk level generator result and ML classifier production into a visual report having a number of sections, including the website information and a summary of information about the data practices in the privacy policy. The following section gives a summary of the risk level for each category. The last section of the report provides associated risk levels, data practices, and the related GDPR article

numbers for each privacy policy segment. Fig. 2 shows an example of a risk report automatically generated from the risk report generator. To make our system usable for non-expert users, we developed a web server to handle the user request to analyze privacy policy text and generate the risk report. Fig. 3 shows the system flow. After passing the request to our back-end risk assessment engine, the engine generates the report and sends it to the front-end to display on the webpage.

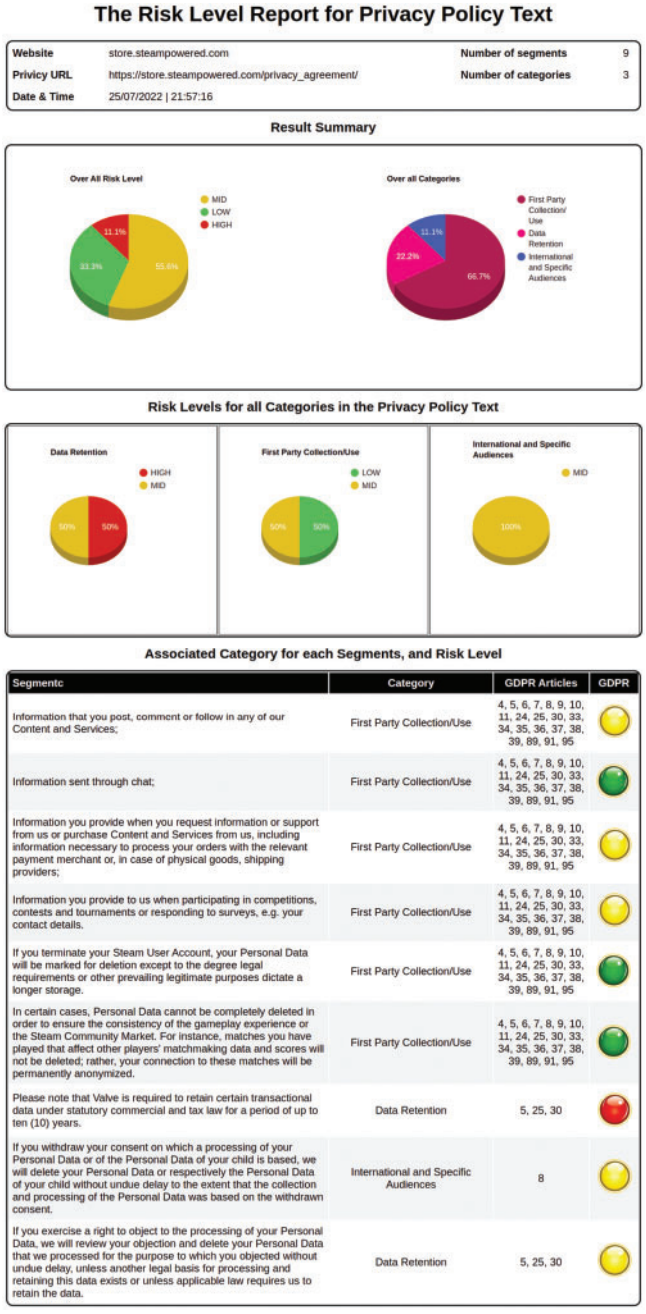

Figure 2: Example of a risk report

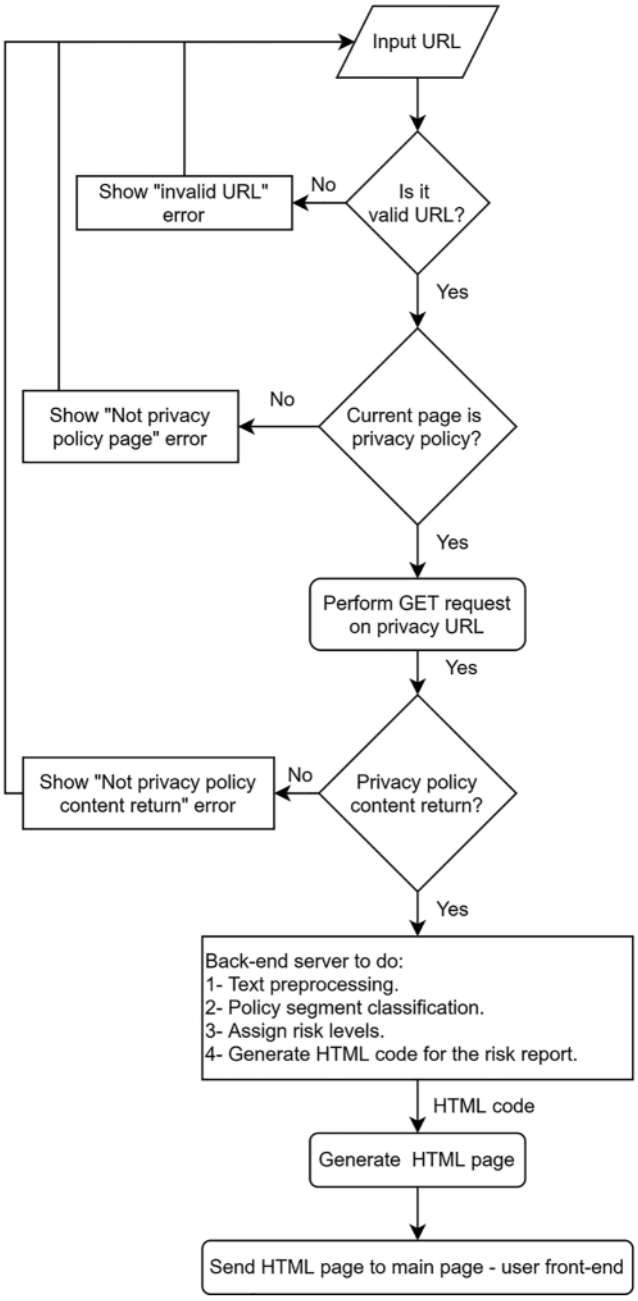

Figure 3: Web service architecture for risk generator

6 Evaluation

To evaluate our approach, we looked at similar approaches that assign a risk level to a given segment/sentence of a privacy policy text. To select these approaches, we looked for projects which had datasets manually annotating the risk level based on the GDPR or which provided clear criteria for assigning a risk level. To the best of our knowledge, PrivacyGuide and Polisis are the only approaches that meet these conditions. The first evaluation was performed with the PrivacyGuide approach.

PrivacyGuide provides a dataset annotating the risk level manually based on the GDPR. Initially, we passed each segment/sentence from the PrivacyGuide dataset to our model, which then assigned the risk level. The final step was to compare our result with the risk level from the PrivacyGuide dataset. Our results matched the Privacy Guide approach rating of risk level at 95.1%. We also achieved a 79% accuracy rate in classifying the paragraph into one of ten categories, whereas the accuracy rate of Privacy Guide is 74%.

The second evaluation was performed with the Polisis approach. This project provided criteria to assign a risk level based on Disconnect icon rules and attribute values from the OPP-115 dataset. Polisis does not provide a dataset for direct use, so we generated a dataset based on the Polisis interpretation and risk assignment description. We then evaluated this approach by performing the same steps as for Privacy Guide. The overall matching between Polisis and our model was 88.72%. Since Polisis does not quite follow GDPR, we investigated which Polisis rules are more closely related to GDPR. Our results show that Data Retention and Children's privacy match at 97% with our model. The lowest match was in the Third-Party Sharing/Collection category, at 72%. Table 8 shows the full results for both evaluation experiments. Our model was the most accurate in classifying the segments into one of ten categories compared with all similar work we covered in related work sections that used traditional ML models and the OPP-115 dataset. Moreover, our model can determine the risk level correctly up to 95% of the time.

Table 8: Matching with other approaches

Approach	% Match
PrivacyGuide	95.1
Polisis-overall	88.72
Polisis-International and Specific Audiences-Children	97.74
Polisis-Data Retention	97.05
Polisis-Personal information type: location	82.69
Polisis-First Party Collection/Use	82.66
Polisis-Third Party Sharing/Collection	72.09

7 Conclusion

Most users agree to privacy policies without reading them because they don't realize the risks to their data and privacy. We have proposed a hybrid approach for a program that can read a policy, classify its data practices into ten categories, assign risk levels to the practices, and provide a visual report for the user. Results show that our approach is not only consistent with GDPR, but also performed better than existing approaches by successfully assigning the risk levels with 95.1% accuracy after assigning the data practice categories with an accuracy rate of 79%. Our approach also mitigates the cost and time incurred with the manual labeling of risk levels as done in existing techniques. In the future, we are planning to explore ways to improve the classification of OPP-115 dataset practices into ten categories and apply ML and DL models to an improved model. We are also working on making the web interface public for general use, so that people can type a privacy policy URL and get the risk level report.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Federal Trade Commission, "Protecting america's consumers," [Online]. Available: https://www.ftc.gov

[2] Organization for Economic Co-operation and Development, "OECD," [Online]. Available: https://www.oecd.org

[3] The European Union, "General data protection regulation," [Online]. Available: https://gdpr.eu/tag/gdpr

[4] State of California Department of Justice, "California consumer privacy Act," [Online]. Available: https://oag.ca.gov/privacy/ccpa

[5] G.-L.-B. Act, "Gramm-leach-bliley Act," 2002. [Online]. Available: https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act.

[6] C. Jensen and C. Potts, "Privacy policies as decision-making tools: An evaluation of online privacy notices," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, Vienna, Austria, pp. 471–478, 2004.

[7] President's Council of Advisors on Science and Technology, "Big data and privacy: A technological perspective," *Report to the President, Executive Office of the President*, May 2014. [Online]. Available: https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf.

[8] A. M. Mcdonald and L. F. Cranor, "The cost of reading privacy policies," *A Journal of Law and Policy for the Information Society*, vol. 4, no. 3, pp. 543–897, 2009.

[9] T. Ermakova, A. Baumann, B. Fabian and H. Krasnova, "Privacy policies and users' trust: Does readability matter?," in *20th Americas Conf. on Information Systems, AMCIS 2014*, Savannah, Georgia, USA, 2014.

[10] K. Litman-Navarro, "We read 150 privacy policies. they were an incomprehensible disaster," in *The New York Times. p. ed*, 12 June 2019. [Online]. Available: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html.

[11] D. Sarne, J. Schler, A. Singer, A. Sela and I. Bar Siman Tov, "Unsupervised topic extraction from privacy policies," in *Companion Proc. of the 2019 World Wide Web Conf.*, San Francisco, USA, pp. 563–568, 2019.

[12] D. A. Audich, R. Dara and B. Nonnecke, "Extracting keyword and keyphrase from online privacy policies," in *Eleventh Int. Conf. on Digital Information Management (ICDIM)*, Porto, Portugal, pp. 127–132, 2016.

[13] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin *et al.,* "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *Proc. of the 27th USENIX Security Symposium*, Baltimore, MD, USA, pp. 531–548, 2018.

[14] R. N. Zaeem, R. L. German and K. S. Barber, "PrivacyCheck: Automatic summarization of privacy policies using data mining," *ACM Transactions on Internet Technology*, vol. 18, no. 4, pp. 1–18, 2018.

[15] A. A. M. Gopinath, S. Wilson and N. Sadeh, "Supervised and unsupervised methods for robust separation of section titles and prose text in web documents," in *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2018*, Brussels, Belgium, pp. 850–855, 2018.

[16] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala *et al.,* "The creation and analysis of a website privacy policy corpus," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, Berlin, Germany, pp. 1330–1340, 2016.

[17] "Art. 12 GDPR-Transparent information, communication and modalities for the exercise of the rights of the data subject-GDPR.eu," [Online]. Available: https://gdpr.eu/article-12-how-controllers-should-provide-personal-data-to-the-subject/

[18] "Art. 13 GDPR-Information to be provided where personal data are collected from the data subject-GDPR.eu," [Online]. Available: https://gdpr.eu/article-13-personal-data-collected/

[19] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang *et al.,* "Have you been properly notified? automatic compliance analysis of privacy policy text with GDPR article 13," in *The Web Conf. 2021-Proc. of the World Wide Web Conf., WWW 2021*, Ljubljana, Slovenia, pp. 2154–2164, 2021.

[20] E. Poplavska, T. B. Norton, S. Wilson and N. Sadeh, "From prescription to description: Mapping the GDPR to a privacy policy corpus annotation scheme," *Frontiers in Artificial Intelligence and Applications*, vol. 334, pp. 243–246, 2020.
[21] "Privacy Icons," [Online]. Available: https://web.archive.org/web/20170709022651/disconnect.me/icons
[22] "Terms of Service; Didn't Read," [Online]. Available: https://tosdr.org/
[23] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *Proc. of the 23rd USENIX Security Symposium*, San Diego, CA, USA, pp. 1–16, 2014.
[24] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto and J. Serna, "PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation," in *CODASPY '18: Eighth ACM Conf. on Data and Application Security and Privacy*, Tempe, AZ, USA, pp. 15–21, 2018.
[25] R. N. Zaeem, S. Anya, A. Issa, J. Nimergood, I. Rogers *et al.,* "PrivacyCheck v2: A tool that recaps privacy policies for you," in *the 29th ACM Int. Conf. on Information and Knowledge Management*, Ireland, pp. 3441–3444, 2020.
[26] A. R. Alshamsan and S. A. Chaudhry, "Machine learning algorithms for privacy policy classification: A comparative study," in *2022 IEEE 2nd Int. Conf. on Software Engineering and Artificial Intelligence (SEAI)*, Xiamen, China, pp. 214–219, 2022.
[27] F. Liu, S. Wilson, P. Story, S. Zimmeck and N. Sadeh, "Towards automatic classification of privacy policy text," *Technical Report CMU-ISR-17-118R*, Carnegie Mellon University, 2018.
[28] A. Kotal, A. Joshi and K. P. Joshi, "The effect of text ambiguity on creating policy knowledge graphs," in *IEEE Int. Conf. Big Data Cloud Comput. (BDCloud 2021)*, New York, NY, USA, pp. 1491–1500, 2021.
[29] Collection definition and meaning, [Online]. Available: https://www.dictionary.com/browse/collection
[30] The 4 types of sentence structure, [Online]. Available: https://www.englishclub.com/grammar/sentence/sentence-structure.htm
[31] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.
[32] M. Kaufmann, "The lokahi prototype: Toward the automatic extraction of entity relationship models from text," in *Proc. of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019)*, Palo Alto, California, USA, 2019.