

Privacy Policy Beautifier: Bringing Privacy Policies Closer to Users

Michalis Kaili and Georgia M. Kapitsaki^a

Department of Computer Science, University of Cyprus, Kallipoleos 75, Nicosia, Cyprus

Keywords: Privacy Policy, Privacy Awareness, User Friendly, GDPR.

Abstract: A plethora of privacy policies are available online, as all websites need to inform the users in detail about the processing of their personal data, especially in the recent years in order to comply with the European General Data Protection Regulation (GDPR). As these texts tend to be very long, understanding their content takes time and is not easy for the majority of users that usually do not spend the time to read the policy texts. In this paper, we present our work on Privacy Policy Beautifier that aims to bring privacy policies closer to the user by highlighting specific parts of the text and presenting the information in different formats: textual with colors, pie chart, word cloud and GDPR terms presence. For the main part of this process, we utilize machine learning techniques. Privacy policy Beautifier has been evaluated via the survey methodology with 90 users and shows potential for assisting in the creation of more user-friendly privacy policy representations.

1 INTRODUCTION


Privacy is considered a fundamental human right (UN, 1948) and in software systems even more emphasis has been placed on its role after the introduction of relevant laws, including the General Data Protection Regulation (GDPR), the Children's Online Privacy Protection Rule (COPPA) and the California Consumer Privacy Act (CCPA). Software systems need to comply with the above laws, whereas they need to make clear how users' data are being collected, stored and processed, including access by third parties.

In Europe, in order to be compliant with the GDPR, privacy policies of these systems are thus, required to contain relevant and specific terms, such as what kind of data are being collected from the user, how the user's data are being used, and which rights the user has, as well as how the user can exercise each right. From the side of the users, being informed about this information is important, as it keeps users privacy aware and assists them in making more educated choices when it comes to what applications they use and which personal information they disclose to these applications (Wagner et al., 2020).

In order to raise users' privacy awareness in an efficient way, it is important to use attractive ways to make the users read the privacy policy of the application, service or website they are intending to use.

Since the data policy texts tend to be long, users rarely read their content and, according to a previous study, are more willing to read shorter texts, making shorter texts helpful in increasing users' reading accuracy and knowledge (Meier et al., 2020). Moreover, in many cases the text of privacy policies is vague (Liu et al., 2016; Lebanoff and Liu, 2018) and may inhibit users' understanding of the privacy policy content. More user friendly ways to present the privacy policy to users are thus required, in order to improve their experience when interacting with web applications, as a means - among other functionalities - to achieve usable privacy.

Having as motivation the above, in this work we present the design, implementation and use of the *Privacy Policy Beautifier* system that aims to provide an easy way for users to read long privacy policy texts and make them more understandable and informative. The creation of *Privacy Policy Beautifier* was thus, motivated by the lack of privacy awareness observed in end users when it comes to using services from the internet (Lebanoff and Liu, 2018; Liu et al., 2016; Pitkänen and Tuunainen, 2012). Despite recent approaches that improve the current situation, this problem is still present (Linden et al., 2018). Previous studies have shown that using colors, 2D tables or images help users increase their privacy awareness more than reading the plain text from the original privacy policy (Soumelidou and Tsohou, 2019), and similar techniques have also been employed in the current work.

^a  <https://orcid.org/0000-0003-3742-7123>

Privacy Policy Beautifier uses machine learning techniques in order to point the user to specific parts of the policy texts, highlighting them according to the category of the text content, and it also presents a pie chart summarizing the categories. The system creates additionally a word cloud of the policy text in the form of a summary, and it indicates GDPR terms that can be found in the text. Existing works examine the usability of privacy policies but do not offer different representations that are created automatically at the same place. *Privacy Policy Beautifier* is making use of colors, tables and pie charts following the results of aforementioned works (Soumelidou and Tsohou, 2019). A prototype web application has been implemented and evaluated via the survey methodology with the participation of 90 users, with most users indicating a positive experience in their interaction with the system.

The rest of the paper is structured as follows. Section 2 provides an overview of related work in the area. Section 3 is dedicated to the process *Privacy Policy Beautifier* uses in order to improve the policy text presentation. Section 4 describes the prototype implementation of the system, demonstrating its use, whereas section 5 provides details on the evaluation process. Section 6 briefly discusses limitations and, finally, section 7 concludes the paper outlining directions of future work.

2 RELATED WORK

Older works in the literature have focused on suggesting new ways of presenting privacy policies to users (Kelley et al., 2009). A survey performed with the participation of 210 Facebook users showed that the majority of active users share a large amount of personal information, whereas they are not aware of how visible this information is to strangers (Pitkänen and Tuunainen, 2012). The survey also showed that privacy policy and terms of use, which need to be accepted by all users so that they can use the service, are largely unknown or not understood.

Another work focuses on giving the users the opportunity to make calculated choices on the distribution of their personal information (Angulo et al., 2012). It presents research results from the PrimeLife project where the PrimeLife Policy Language (PPL) was created and evaluated with users via the use of the “Send Data?” browser extension prototype that presents to the user fundamental elements of a service provider’s privacy policy in an easy to understand and user-friendly way, showing both advantages and some challenges in its use.

Regarding GDPR compliance and privacy policies, a tool that analyzes the degree of privacy policies indicating GDPR terms, CompLicy, has been recently introduced (Vanezi et al., 2021). The aim of CompLicy is to show how well a privacy policy text integrates the rules and principles of GDPR. A parser is used to extract the required text from the webpage that contains the privacy policy. The text is then analysed and processed in combination with a list of keywords and phrases relevant to GDPR. A score is given to the privacy policy to show how well it covers all necessary points of the GDPR, whereas a more detailed analysis shows which points have been included and what is missing from the privacy policy.

In order to examine the impact of GDPR on privacy policies, a corpus of 6,278 unique English-language privacy policies from both inside and outside the European Union was created and was then compared to their versions before the introduction of GDPR (Linden et al., 2018). According to this work, GDPR has lead to major changes in the privacy policy landscape with most changes being in EU-based websites. It was also observed that privacy policies have become significantly longer in length, probably to cover and satisfy the new regulations. Despite being more extensive, the new privacy policies have also improved upon their visual representation making them more appealing to the end users. It is also noted that previous regulations changed the privacy policy landscape with more websites adopting or changing their privacy policies as well, as some of them are becoming more extensive and descriptive. But that always came at the cost of readability and clarity of the privacy policy. Despite the improvement in visual representation, policy texts remain long for users to be able to read without devoting a large amount of time.

The study by Galle Mattias et al. focuses on the advantages and disadvantages for the need of a dataset of privacy policies, annotated with GDPR-specific elements (Gallé et al., 2019). The authors revise existing and related datasets to see how they could be modified or changed in order to make it possible for them to be used in various machine learning techniques. They highlight that with the introduction of the GDPR, Natural Language Processing (NLP) techniques can be very beneficial especially to small businesses and enterprises that are trying to be compliant with GDPR. The paper concludes that the current datasets should be revised taking into consideration four aspects related to GDPR: impact of new GDPR elements, impact of multi-linguality, impact of domain shift due to the type of companies, and impact of domain shift due to the adaptation to the GDPR.

Although more work is needed towards datasets

of privacy policies, the most widely used assistance in understanding privacy policies has been offered by the OPP-115 dataset that is used also in the framework of the current work (Wilson et al., 2016). This work presents the creation process (e.g. privacy policy selection) and the structure of the dataset, its composition and initial experimentation.

Concerning the visualisation and presentation of privacy policies there have been attempts that improve the visualisation using manual or automatic techniques. An existing study proposes new visualisation techniques for privacy policies instead of the traditional textual representation, but with an emphasis on how each technique affects the users and their privacy awareness (Soumelidou and Tsohou, 2019). Three different visualizations were used: Unchanged policy text, Tag clouds (WordBridge) and Document Cards. No automatic technique was used but the privacy policy of Instagram was used and was transformed in the aforementioned representations. The evaluation performed with students showed higher policy awareness using the two new visualization techniques compared to the traditional textual representation. This work strengthens the motivation of the current work in using different visualizations, but we focus instead on automating the approach.

When it comes to more automated means to analyze, process and transform the privacy policy to make it easier for the average user to understand, deep learning and graphs for representing the necessary information are proposed in Polisis that utilises the aforementioned OPP-115 dataset (Harkous et al., 2018). Together with Polisis, Pribot was also created in the same work in order to assist in answering structured and free form queries concerning the policy (Harkous et al., 2018). Privacy policy icons from privacy policies are assigned automatically with an accuracy of 88.4%, whereas free form question answering provides a correct response in the top three results for 82% of the test questions.

An approach that relies on data mining to create summaries of privacy policies to give the user the general idea in a smaller, easier to read, and understandable chunk and color coded symbols is found in PrivacyCheck (Zaem et al., 2018). PrivacyCheck is a browser extension that presents to the user a summary of the contents of the privacy policy. It relies on classification techniques analyzing the policy text in terms of user control and GDPR content.

Other researchers have tried using unsupervised methods, either to see whether word embedding specifically for privacy policies can help other researchers in their endeavor to automate this process (Kumar et al., 2019), or to extract topics from

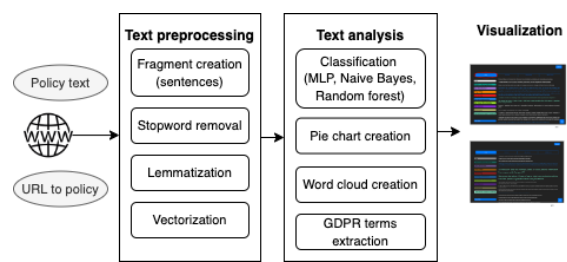


Figure 1: Processing steps of *Privacy Policy Beautifier*.

privacy policies and unveil new ones that supervised methods might have missed (Sarne et al., 2019).

In relation to previous works, *Privacy Policy Beautifier* falls within automated approaches and offers an environment with different representation options. Although the accuracy achieved in the classification process it employs is lower than that of some previous works (e.g. Polisis), its main advantages compared to previous works are that: 1) it combines a number of techniques (i.e. text highlighting, pie chart, word cloud, GDPR terms presence) that allow users to inspect the privacy policy from different perspectives in a fast way, 2) it has been evaluated online with internet users.

3 PRIVACY POLICY BEAUTIFIER APPROACH

Privacy Policy Beautifier aims to provide users a more user friendly way in the representation of privacy policies available on the web. For this purpose, it provides different representations of the various parts of the text, specifically 1) by highlighting parts of the text that refer to specific areas of interest, i.e. categories, such as data category, data retention, 2) by presenting a pie chart with the encountered categories, 3) by presenting a word cloud of the original policy text, and 4) by indicating which GDPR relevant terms are present in the text. In order to perform the text highlighting, which is the main functionality of the proposed system, supervised machine learning techniques are employed, by applying classification on the policy content. The different parts of the text, after being separated and labeled, are presented to the user in a highlighted fashion so that the user can more quickly navigate to specific parts of the policy text. This gives the user the opportunity to read the information he/she wants without spending more time looking through the whole text to find a specific aspect of the policy he/she is interested in. The steps used by *Privacy Policy Beautifier* are depicted in Figure 1 and are described in more detail in the remaining of the section.

3.1 Text Highlighting

Text highlighting is the main functionality of *Privacy Policy Beautifier* which utilises supervised machine learning to create a model in order to classify the different parts of the privacy policy text. These classified segments are then presented to the user with different colors to make it easier and faster for the user to find information he/she is most interested in.

The training and testing of the classifier has been performed using the OPP-115 dataset that has been widely used in previous works (Wilson et al., 2016). The text is highlighted using the following 10 categories indicated in the OPP-115 dataset:

- *First Party Collection/Use*: how and why data are collected by the service provider.
- *Third Party Sharing/Collection*: how data are collected or shared with third parties.
- *User Choice/Control*: choices and controls for users.
- *User Access, Edit and Deletion*: if and how users can perform the above actions.
- *Data Retention*: duration of storing data.
- *Data Security*: security applied on data.
- *Policy Change*: if and how users are informed about changes.
- *Do not Track*: how is “do not track” applied.
- *International and Specific Audiences*: practices applicable to specific user groups.
- *Other*: text not covered in other categories (also covering introductory text of the policy).

The classifier, given a sentence from a privacy policy text, positions the sentence into one of the 10 categories indicated above.

Text Preprocessing. In order to prepare the text, data not useful for the classification process were removed in this step. Initially, the policy text was broken down into sentences. English stopwords were then removed. The list of stopwords used contains the standard stopwords from the NLTK (Natural Language Toolkit) library of Python, which includes 127 words. Lemmatization was applied on the terms as a next step. Note that the use of stemming was also examined but the experiments with different classifiers as described next provided better results with lemmatization. Finally, the text was converted to vector representation. Different techniques were also employed in this step and they were compared to achieve the highest accuracy, i.e. term frequency-inverse document frequency (TF-IDF), bag-of-words, with TF-IDF adopted as the final technique.

Table 1: Comparison of classification algorithms for text highlighting.

	MPL	Naive Bayes	Random forest
Accuracy	50%	60%	74%

Training and Algorithm Comparison. The following classification algorithms that are widely employed in text categorization were compared (Dhar et al., 2021):

- Multilayer perceptron (MLP) (Gardner and Dorling, 1998): MLP is a feedforward neural network which is composed of multiple layers of perceptrons.
- Naive Bayes classifier (Leung, 2007): it is based on the Bayes Theorem with an assumption of independence among predictors.
- Random forest classifier (Liaw et al., 2002): it is a supervised learning algorithm that fits or trains a set of decision tree classifiers.

The results of the training process in terms of classifier accuracy that measures the number of correct predictions in relevance to the total number of predictions are listed in Table 1. Note that for the case of MLP various values were examined for the number of hidden layers and the number of perceptrons contained in each layer and different activation functions and solvers were tested (with varying batch sizes and learning rates), but the accuracy did not improve. For the case of Naive Bayes, different vectorization techniques were used, as well as different feature extraction methods. In the case of the random forest classifier that achieved the highest accuracy, lemmatization as part of the preprocessing steps provided an accuracy of 74% compared to 70% with the use of stemming.

Since the Random forest classifier has the highest accuracy, it was adopted in the implementation of the web platform of *Privacy Policy Beautifier*. We also employed recall, precision and F1-score in order to examine how random forest behaves in the different categories available in the OPP-115 dataset. Recall is the fraction of relevant documents or items that are successfully retrieved. Precision is the number of correctly classified items given by the classifier. The F1-score combines precision and recall as follows:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

These three metrics were not used in order to select the algorithm, but they assisted in finding potential issues in the training phase that were then addressed by using different parameters in the algorithm or by adjusting the preprocessing steps. Precision, re-

Table 2: Precision, recall and f1-score values for the random forest classifier.

	Precision	Recall	F1-score
First Party Collection/Use	0.73	0.81	0.77
Third Party Sharing/Collection	0.79	0.69	0.74
User Choice/Control	0.83	0.58	0.68
User Access, Edit and Deletion	0.76	0.53	0.63
Data Retention	0.67	0.33	0.44
Data Security	0.77	0.80	0.79
Policy Change	0.88	0.78	0.82
Do Not Track	1.00	0.83	0.91
International and Specific Audiences	0.60	0.45	0.51
Other	0.70	0.80	0.75

call and F1-score values for the random forest classifier are listed in Table 2. The table depicts the values per classification category. Precision ranges between 0.67 and 1.00, recall between 0.33 and 0.83, and F1-score between 0.44 and 0.91, with the lowest values observed in the “Data Retention” category and the highest in the “Do Not Track.”

3.2 Word Cloud Summary

We use tag clouds as this is a text visualization technique employed widely in previous works not only related to privacy policies (Kim et al., 2011; Paskali et al., 2021). The less frequent a word is the smaller it appears in the word cloud. This representation was included to give the user a general overview of what is being said in the privacy policy and how much of each term they are expected to see without having to go through the entire text. The original policy text was used ignoring stopwords.

3.3 GDPR-relevant Terms

For this part, the set of GDPR terms available in a previous work consisting of 89 terms was utilized (Vanezi et al., 2021). The terms cover the following GDPR categories:

- *Lawfulness of Processing* (example terms: Lawfulness of Processing, Consent, Contract, Right to Withdraw Consent, Withdraw consent)
- *Right to Restriction of Processing* (example terms: Restriction of Processing, Restrict Your data, Right to Restrict, Right to demand processing restrictions, Right to restriction of processing)
- *Right of Access by the Data Subject* (example terms: Right Of Access, Right To Access, Access Personal Data, Access Your Personal Data, Access your data)
- *Right to Data Portability* (example terms: Right to Data Portability, Right to Transmit Those Data, Transmit Your Personal Data, Request the transfer of your personal data, Request the transfer data)

- *Right to Rectification* (example terms: Right to have incomplete personal data, Right to Request Proper Rectification, The Right to Correct and Update, Rectify Your Data, Update or Correct your Information)
- *Right to Object* (example terms: Right to Object at any time to Processing of Personal data, Right to Object to Processing, Right to Object at any time to Processing, Processing Objection, Object to processing)
- *Right to Erasure* (example terms: Right of Erasure, Right to Request Deletion, Right To be Forgotten, Request erasure, Erase your Personal data)

Note that in relevance to text highlighting, different categories are used, as this part of the analysis is specific to GDPR’s provisions and rights that are relevant to web platforms, and are therefore more specific to the user and the rights the user can exercise. These rights are only partially present in the OPP-115 dataset categories and there is a small overlap that can be found in “User Choice/Control” and “User Access, Edit and Deletion” categories.

A simple approach was followed in this step, where the exact terms that cover the indication of the same GDPR right with different wording are searched in the privacy policy text. The user is informed about the presence of the term as an indicator of compliance of the policy text, and thus of the respective application, to GDPR.

4 USE DEMONSTRATION

Implementation. Python was the main programming language used with the scikit-learn library employed primarily for the classification process, along with additional libraries mentioned earlier, such as NLTK used for the stopword removal. Concerning the web system implementation, the front end was created using the Flask² web framework, along with the use of HTML/CSS, JavaScript and Bootstrap³. Moreover, libraries such as Google Charts and WordCloud2 were used for the visual representation of information.

Regarding the use of the system, the user has two options in order to add a new privacy policy for analysis. He/she can either point to a URL where the text is located, or he/she can copy and paste the respective text as depicted in Figure 2. After the execution of the steps described in the previous section the results are displayed to the user in different tabs of the User Interface for the four supported representations:

²<https://flask.palletsprojects.com/>

³<https://getbootstrap.com/>

Figure 2: Privacy policy selection in *Privacy Policy Beautifier*.

- Regarding the text highlighting results, the classified segments given by the classifier are color coded and dynamically inserted in such a way as not to change the original structure of the privacy policy, as shown in Figure 3. By choosing a specific category from the options on the left, the user is directed to the respective parts of the text that contain text of that category. This gives the user the opportunity to focus on that specific aspect of the policy. The category filtering can be activated and deactivated by the user several times, when interacting with *Privacy Policy Beautifier*. In order to show to the user on which categories the privacy policy puts more emphasis and whether or not a certain category is present, the pie chart is also used as summary indicating the presence of the different parts as depicted in Figure 4. Finally, the word cloud representation of the terms found in the policy and the GDPR terms presence are also visible to the user in their respective tabs (Figures 5 and 6 respectively). The presence of each GDPR term is indicated in green color and the absence in red. In the provided example, only a small number of GDPR terms are present in the text.

5.1 Questionnaire Design

				Subs
Policy	Privacy	WordCloud	GDPR Terms	
Other	<p>Privacy Policy Sci News.com's Sci News.com is committed to protecting and respecting your privacy.</p> <p>To better inform you of our policy concerning user privacy, we have adopted the following terms.</p> <p>Please note that these terms are subject to change, and any such changes will be indicated on this page.</p> <p>Information that Sci News.com may collect Online Sci News.com may collect and process the following data about you - information that you provide by filling in forms on our site, including names, e-mail and address; we may also ask you for information for other purposes, for example when you make a request and problem with our site. If you contact us, we will keep a record of correspondence - details of your visits to our site are included, but not limited to, traffic, location data, weblogs and other communication data.</p> <p>Sci News.com does not knowingly collect or solicit personal information from anyone under the age of 13.</p> <p>We assume that minors 13 years of age and older have received permission from their parents or guardians before using this website.</p> <p>Parents or guardians may contact us at privacy@sci-news.com with questions or concerns about our privacy policy.</p> <p>Use of Cookies Sci News.com uses "cookie" technology.</p> <p>A cookie is a small amount of data, which often includes a unique identifier that is sent to your computer or mobile phone (browser from a web browser computer) and is stored on your device hard drive.</p> <p>A website can send its own cookie to your browser if your browser's preferences allow it, but your browser only permits a website to access the cookies it has already sent to you, not the cookies sent to you by other websites.</p> <p>Many websites do this whenever a user visits their website in order to track online traffic flows.</p>			
Third Party Sharing/Collection				
User Choice/Control				
Policy Change				
Data Security				
Third Party Collection/Use				
User Access, Edit and Deletion				
Data Retention				
Do Not Track				
International and Specific Audiences				
User Info				

Category percentage

Category	Percentage
Other	71.4%
Third Party Sharing/Collection	7.1%
First Party Collection/Use	7.1%
User Choice/Control	7.1%
International and Specific Audiences	7.1%
Policy	7.1%

[illegible]

Policy	PicChart	WordCloud	GDPR Terms
Lawfulness of Processing			
Lawfulness of Processing			NO
Consent			YES
Contract			NO
right to withdraw consent			NO
Withdraw consent			NO
Right of Erasure			
Right of Erasure			NO
The Right to Request Deletion			NO
Right To be Forgotten			NO
Erase your information			NO
Request erasure of your personal data			NO
Erase the Personal data			NO

59

ity, privacy awareness. A dedicated questionnaire was created for this purpose and was then distributed to various users via email communication. Contacts of the authors were used for the email communication and they were asked to forward the questionnaire to more prospective participants from their contacts.

Access to the *Privacy Policy Beautifier* was also provided to the participants. The users were asked to first use the system, using either some example privacy policies indicated or one of their choice (some examples provided were: SciNews⁴, IHS Markit⁵, Biogen⁶, Veeam⁷). The participants were then asked to answer the questionnaire to document their experience. The questionnaire is available online⁸. All participants provided their consent after being informed about the questionnaire's purpose and the relevant use of the collected data. In the first part, demographic information about the participants was gathered (i.e. age, gender, level of education, educational background). The questionnaire contained questions about users experiences with reading privacy policies in the past. Finally, the participants were asked to provide their opinion on their interaction with *Privacy Policy Beautifier*, including some aspects of usability, such as how easy it was to use *Privacy Policy Beautifier* and whether they would use it again.

5.2 Analysis of Results

In total 90 individuals participated in the survey. The participants are primarily from Cyprus, as mainly local contacts were reached. The age of the participants ranged between 18 and above 60 years old: most participants were 18-24 years old (41.1%), followed with the 25-29 group (18.9%), the 40-49 and 50-60 age groups (both 14.4%), and the 30-39 age group (8.9%). Most participants were male (55.6%), whereas some did not disclose the gender information (4.4%). The majority of participants have limited technical experience (41.1%), some did not have any technical experience (33.3%) and the remaining 25.6% did.

Only 28.9% of participants indicated that they have read the privacy policy of a website in the past (54.4% indicated that they have partially read a policy). This indicates that some users are spending time and effort to go through privacy policies, which is a good sign that they are interested in finding information concerning privacy. Out of the 75 participants that have read a policy, either fully or partially,

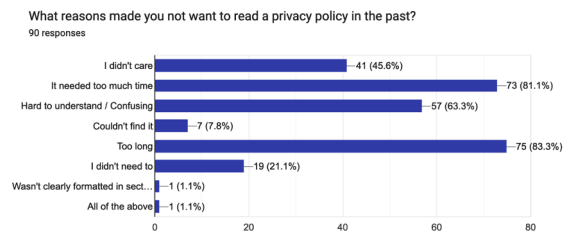


Figure 7: Reasons that made participants not read a policy text in the past.

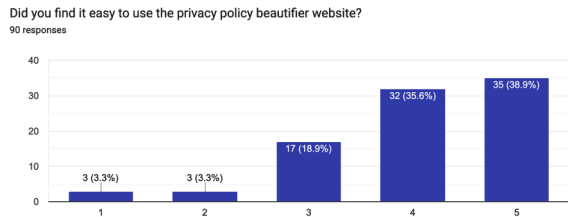


Figure 8: Participants view of how easy it was to use *Privacy Policy Beautifier*.

most did not find the process enjoyable as expected (61.3%) and had some hard time finding information they needed (54.7%). Out of these 75 participants, 16% indicated that they would not read a privacy policy again in the future. The reasons mentioned for this were that privacy policies are too long, too complicated, or contain too much information or are a waste of time. Participants were asked to provide free text to indicate these reasons.

The above reasons are aligned with the top 3 reasons participants indicated as reasons that made them not want to read a privacy policy in the past: too long, time consuming or too hard to understand/confusing (Figure 7). Unfortunately a large number of people indicated a lack of interest in increasing their privacy awareness ("I didn't care" answer).

Regarding the interaction with *Privacy Policy Beautifier*, participants were asked whether they found the web system easy to use and whether or not it made the privacy policy easier to read. 5-point Likert scale was used in these questions (ranging from 1-very hard to 5-very easy in the first question, and from 1-strongly disagree to 5-strongly agree in the second question). The results are visible in Figure 8 and in Figure 9 respectively, where most participants indicated that they could easily (4) or very easily (5) use the web platform and were positive about the experience in terms of privacy policy readability: most participants agreed (4) or strongly agreed (5). Most users were also positive about using *Privacy Policy Beautifier* in the future, as shown in Figure 10.

In order to better understand how the profile of the users affects their privacy awareness and their inter-

⁴<http://www.sci-news.com/privacy-policy.html>

⁵<http://www.ihsmarkit.com/Legal/privacy-policy.html>

⁶http://www.biogen.com/en_us/privacy-policy.html

⁷<http://www.veeam.com/privacy-policy.html>

⁸<https://forms.gle/URefBrGwGXD5DwWi9>

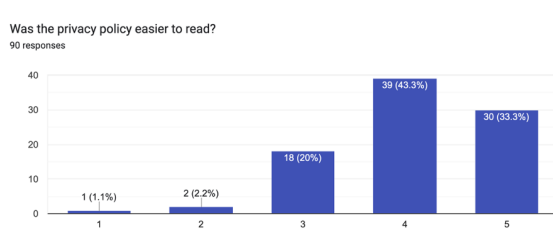


Figure 9: Participants view of how *Privacy Policy Beautifier* improves the policy readability.

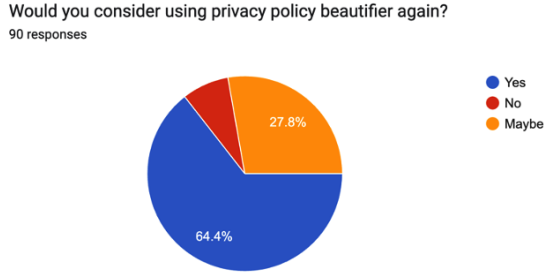


Figure 10: Participants view of whether they would use *Privacy Policy Beautifier* again.

action with *Privacy Policy Beautifier*, we ran a number of statistical tests. Initially, we examined whether the technical background of the users affects how easy they find the system, whether the policy is easier to read via the proposed system and whether they would use *Privacy Policy Beautifier* in the future (that was another question included in the questionnaire). We ran one-way ANOVA test for this purpose but no statistically significant differences were observed for the above parameters indicating that the system is easy to use regardless of the technical background of the users. The same observation was made, by running one-way ANOVA to examine whether having read a policy in the past affects the above parameters. Regarding using the system in the future, most users were positive (64.4% said they would use it and 27.8% said that they would probably use it).

Regarding the representation type, where participants could choose one or more preferred representations, most users preferred the textual representation with text highlighting (55.6%), followed by the pie chart (45.6%) and the GDPR terms (42.2%), whereas the word cloud representation was not that popular among the participants (only 15.6% indicated it in their answers), as depicted in Figure 11.

Table 3: Preference for textual representation considering participants' technical expertise.

Technical expert	Text highlighting	Pie chart
Yes	78.3%	26.1%
No	43.3%	60%
I know some stuff	51.4%	45.9%

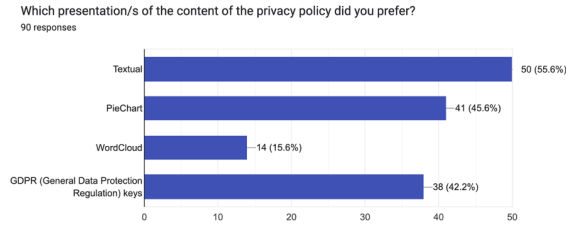


Figure 11: User preferences for the different representations.

We examined whether the technical background of participants affects the representation preferences of the privacy policy. One-way ANOVA revealed that no statistically significant results appear for the word cloud and the GDPR terms representation, but there is a statistically significant difference concerning the textual ($p = 0.031$, $F = 3.602$) and the pie chart representations ($p = 0.049$, $F = 3.130$), as shown in Table 3. Technical experts show larger preference in the textual representation type compared to non-experts and participants with limited technical expertise but less preference in the pie chart representation than the two other user groups. Overall, users that are technical experts prefer the textual and the GDPR representations, 18 out of 23 and 13 out of 23 respectively, but they do not prefer the pie chart representation (3 out of 23). In contrast, non-technical users seem to prefer the pie chart representation (18 out of 30).

Finally, we examined whether the age of the participants affected any of the above parameters. A statistically significant result was found only for whether it was easier to read the privacy policy via the use of *Privacy Policy Beautifier* ($p = 0.032$, $F = 2.585$). Participants of younger ages were more positive that the proposed system helped them understand privacy policies better (Table 4), although we believe that this observation may not be directly related with the use of the system but be a more general observation regarding the interaction with any web system when it comes to users' age group.

6 LIMITATIONS

Concerning the accuracy of the text highlighting approach, the most important limitation is the small size of the training data and the fact that the data do not have a wide range of diversity. Some specific aspects of the classifiers may have affected the accuracy of each one, such as the zero frequency problem in the case of Naive Bayes, where the algorithm assigns a zero probability to a categorical variable whose category in the test dataset was not available in the train-

Table 4: Users that found the privacy policy easier to understand per age group.

Age group	# users	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
18-24	36	52.8%	33.3%	11.1%	2.8%	0%
25-29	17	29.4%	47.1%	17.6%	5.9%	0%
30-39	8	12.5%	50%	25%	0%	12.5%
40-49	13	7.7%	69.2%	23.1%	0%	0%
50-59	13	30.8%	38.5%	30.8%	0%	0%
≥60	2	0%	0%	100%	0%	0%

ing dataset, may have lead to a lower accuracy in the results.

The classifier shows lower accuracy than Polis (Harkous et al., 2018), which has an average accuracy of 88.4% but it is on the same level as other approaches such as PrivacyCheck (Zaeem et al., 2018), which has an accuracy of 40%-73%. Future work will examine whether the use of unsupervised techniques or a combination of supervised and unsupervised techniques used in previous works can improve the accuracy (Harkous et al., 2018; Sarne et al., 2019).

Regarding the user study, it is affected by *external validity*, referring to the extent we can generalize our findings. Although our user sample included users of various backgrounds (i.e. ages and expertise), a larger sample may provide slightly different observations.

7 CONCLUSIONS

In this paper, we have presented our work on more user friendly representations of the text of privacy policies via *Privacy Policy Beautifier*, where policies are presented to users in different ways: textual with text highlighting, in the form of a pie chart, as a word cloud and as a table with indications of the presence of GDPR terms. The classification accuracy of the proposed classifier shows promising results (74%) that can be further improved, whereas the user study showed that users value the different representations with many users having a positive interaction with different representations. Future work will examine whether the use of unsupervised techniques or a combination of supervised and unsupervised techniques can improve the classification accuracy. As part of future work, we also intend to enhance *Privacy Policy Beautifier* by adding a summarization of the text, and by considering the addition of more representations that have been studied in previous works, while studying their effect on the user experience (Soumelidou and Tsohou, 2019).

REFERENCES

- Angulo, J., Fischer-Hübner, S., Wästlund, E., and Pulls, T. (2012). Towards usable privacy policy display and management. *Information Management & Computer Security*.
- Dhar, A., Mukherjee, H., Dash, N. S., and Roy, K. (2021). Text categorization: past and present. *Artificial Intelligence Review*, 54(4):3007–3054.
- Gallé, M., Christofi, A., and Elshahar, H. (2019). The case for a gdpr-specific annotated dataset of privacy policies. In *AAAI Symposium on Privacy-Enhancing AI and HLT Technologies*.
- Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.
- Harkous, H., Fawaz, K., Lebre, R., Schaub, F., Shin, K. G., and Aberer, K. (2018). Polis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} security symposium ({USENIX} security 18)*, pages 531–548.
- Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. (2009). A “nutrition label” for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12.
- Kim, K., Ko, S., Elmqvist, N., and Ebert, D. S. (2011). Wordbridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–8. IEEE.
- Kumar, V. B., Ravichander, A., Story, P., and Sadeh, N. (2019). Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.
- Lebanoff, L. and Liu, F. (2018). Automatic detection of vague words and sentences in privacy policies. *arXiv preprint arXiv:1808.06219*.
- Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007:123–156.
- Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.
- Linden, T., Khandelwal, R., Harkous, H., and Fawaz, K. (2018). The privacy policy landscape after the gdpr. *arXiv preprint arXiv:1809.08396*.

- Liu, F., Fella, N. L., and Liao, K. (2016). Modeling language vagueness in privacy policies using deep neural networks. In *2016 AAAI Fall Symposium Series*.
- Meier, Y., Schäwel, J., and Krämer, N. C. (2020). The shorter the better? effects of privacy policy length on online privacy decision-making. *Media and Communication*, 8(2):291–301.
- Paskali, L., Ivanovic, L., Kapitsaki, G., Ivanovic, D., Surla, B. D., and Surla, D. (2021). Personalization of search results representation of a digital library. *Information Technology and Libraries*, 40(1).
- Pitkänen, O. and Tuunainen, V. K. (2012). Disclosing personal data socially—an empirical study on facebook users’ privacy awareness. *Journal of Information Privacy and Security*, 8(1):3–29.
- Sarne, D., Schler, J., Singer, A., Sela, A., and Bar Siman Tov, I. (2019). Unsupervised topic extraction from privacy policies. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 563–568.
- Soumelidou, A. and Tsohou, A. (2019). Effects of privacy policy visualization on users’ information privacy awareness level: The case of instagram. *Information Technology & People*.
- UN (1948). Universal declaration of human rights. *United Nations General Assembly*, 302(2):14–25.
- Vanezi, E., Zampa, G., Mettouris, C., Yeratziotis, A., and Papadopoulos, G. A. (2021). Complicity: Evaluating the gdpr alignment of privacy policies-a study on web platforms. In *International Conference on Research Challenges in Information Science*, pages 152–168. Springer.
- Wagner, C., Trenz, M., and Veit, D. (2020). How do habit and privacy awareness shape privacy decisions?
- Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., et al. (2016). 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.
- Zaeem, R. N., German, R. L., and Barber, K. S. (2018). Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.