

Dissertation

PRIVACY POLICY BEAUTIFIER

Michalis Kaili

UNIVERSITY OF CYPRUS



DEPARTMENT OF COMPUTER SCIENCE

December 2020

UNIVERSITY OF CYPRUS DEPARTMENT OF COMPUTER SCIENCE

Privacy Policy Beautifier

Michalis Kaili

Supervisor

Dr. Georgia Kapitsaki

The Individual Diploma Thesis was submitted towards partially meeting the requirements for obtaining the degree of Computer Science of the Department of Computer Science of the University of Cyprus

December 2020

Acknowledgements

This thesis would not have been possible without the help and support of certain individuals.

I would first like to thank my supervisor, Professor Kapitsaki Georgia, who has provided me with all the help and guidance that I required to make this thesis possible. Your insightful feedback has motivated me to improve my thinking and bring my work to a level I never thought I could achieve.

Additionally, I would like to thank my family who have always stood by my side and encouraged me to pursue my goals and provided me with all the necessities I needed to do so. More specifically, my mother and sister who helped shape who I am as a person and with which I could never imagine my life without.

Finally, I would like to thank my dear friends, Aris, Joanna, Nicholas, Olympia and Panayiotis who provided me with joy, happiness, love and more support that they will ever know or admit.

Abstract

In this individual thesis the goal is to help people raise their privacy awareness by encouraging them to read privacy policies they encounter. This is done with a web application that tries to make privacy policies more user friendly by separating its contents, color coding it and providing additional information with charts or 2D tables.

A general description of the meaning and importance of privacy will be given as well as the dangers that threaten it in our modern society. Additionally, the GDPR will be looked at due to its strong impact in protecting the privacy of EU citizens.

Moreover, articles, studies and tools that have taken on the task of improving or analysing different aspects of privacy policies will be thoroughly examined. The knowledge contained in these articles has given great insight and contributed greatly in the creation of this thesis.

Next, the Privacy Policy Beautifier that was developed in the context of this thesis will be presented and analysed. The inner workings of the web application will be fully described. This includes the frontend, backend and the classifier that was created and trained. Furthermore, the tools that were used will be mentioned followed by the reasons that lead to those choices. Additionally, explanations are given on why some design decisions were made and how it was hoped they would affect the outcome of the project.

The system went through constant evaluation during its development phase to make sure its performance was adequate. But the evaluation continued after the system was deployed to make sure that the users were not having issues while using it.

Finally, a conclusion was reached concerning the effectiveness of this web application. It was observed that users had a more pleasant experience while using the Privacy Policy Beautifier instead of the original privacy policy. Additionally, a discussion on the need and usability of this web application takes place with some suggestions for future works both on the improvement of this project and its usefulness in future projects.

Contents

Chapter 1	1
Introduction	1
1.1 Motivation	1
1.2 Concept	2
1.3 Methodology	3
1.4 Chapter Outline	5
Chapter 2	6
Background	6
2.1 Personal Data and Privacy	6
2.2 Privacy Threatened	9
2.3 General Data Protection Regulation (GDPR)	11
Chapter 3	14
Related Work	14
3.1 Introduction	15
3.2 Existing works on privacy policies	15
3.2.1 Disclosing Personal Data Socially - An Empirical Study on Facebook Users' Privacy Awareness	16
3.3 Existing works on privacy policies and GDPR	18
3.3.1 The Privacy Policy Landscape After the GDPR	19
3.3.2 The Case for a GDPR-specific Annotated Dataset of Privacy Policies	20
3.3.3 Evaluating privacy policies on web platforms based on the GDPR	22
3.4 Research on the topic of improving the visualization of privacy policies	22
3.4.1 Effects on privacy policy visualization on users' information privacy awareness level (instagram)	24
3.4.2 Polisis Automated Analysis and Presentation of Privacy Policies Using Deep Learning	26
3.4.3 PrivacyCheck Automatic Summarization of Privacy Policies Using Data Mining	27
3.4.4 Quantifying the Effect of In-Domain Distributed Word Representations: A Study of Privacy Policies	29
3.4.5 Towards usable privacy policy display and management	30
3.4.6 Unsupervised Topic Extraction from Privacy Policies	32

Chapter 4	34
Privacy Policy Beautifier and how it works	34
4.1 Introduction	34
4.2 Training The Classifier	36
4.2.1 Background	37
4.2.2 Training	37
4.2.3 Previous attempts	43
4.2.4 Final Classifier	44
4.3 Privacy Policy Beautifier	45
Chapter 5	53
Platform Evaluation	53
5.1 Introduction	53
5.2 Classifier Evaluation	54
5.3 User Evaluation	56
5.3.1 The questionnaire	56
5.3.2 Evaluation Results	58
Chapter 6	69
Conclusion, Discussion and Future Work	69
6.1 Introduction	69
6.2 Conclusion	69
6.3 Discussion	71
6.4 Future Work	72
Bibliography	73
Appendix A	A-1
Questionnaire	A-1
Questionnaire Answers	A-8

Chapter 1

Introduction

1.1 Motivation	1
1.2 Concept	2
1.3 Methodology	3
1.4 Chapter Outline	5

1.1 Motivation

It is apparent that in our day and age people have started to transition into a more digital lifestyle. This is caused by the rapid advances in technology which provide many life improvements or make people's desires more accessible. This new lifestyle means that people now use electronic devices like smartphones, tablets, smartwatches, and more as part of their life and daily routine. This means that most people have made the regular use of online services (amazon, ebay, spotify, etc), social media networks (facebook, instagram, viber, whatsapp, snapchat, reddit, etc) and much more as part of their life. Most of these services require the users to insert and disclose lots of personal information, in addition to collecting data on their own about the user's activity, likes/dislikes, friends, etc. This massive collection and use of each individual's personal information has become a large concern in recent years, not only for the potential of companies taking

advantage of this information but for data leaks that have been observed many times in the past.

This problem as mentioned above has been on the mind of people and governing bodies for some time now. Some attempts have been made to minimize this issue or to inform the users in order for them to be more aware of what data they are disclosing and how that data is being used, but the problem still endures. But not all hope is lost just yet as some of the attempts made to tackle this issue have had a major impact on the world of privacy policies.

Privacy Policies, especially after the introduction of the GDPR [23], are required to contain very important subjects like what data is collected from the user, how the user's data is being used, what rights the user has, and much more. Knowing this information will help users raise their privacy awareness and make more educated choices when it comes to what services they use and what personal information they disclose.

The problem mentioned above is the motivation behind this project, which aims to help users raise their privacy awareness by making it more appealing for them to read the privacy policies of websites or applications they use. The combination of the GDPR, which forces privacy policies to be more transparent and the Privacy Policy Beautifier, which will encourage users to read them, will hopefully help users raise their privacy awareness.

1.2 Concept

To achieve the desired effect of raising the users' privacy awareness it is necessary to make the user read the privacy policy of the site/application they are intending to use. This task is not easy because users have not had many pleasant experiences in the past when it comes to privacy policies. This is due to privacy policies tending to be massive walls of text that require a large amount of time and effort to read. So to make the user more inclined to read a privacy policy

it has to either be broken down into its different parts, which the user can then browse to find what they are interested in, or to summarize the entire text.

It was decided that breaking the privacy policy into its parts was a better approach because summarizing such an important document meant there was a huge risk of losing important information. The different parts, after being separated and labeled, will then be presented to the user in an easy to understand way such that the user can then find what they are looking for and be directed to it. In this way the user will be able to locate and read the information they want without spending too much time looking through unwanted information. Additionally, more data will be extracted from the privacy policy and presented to the user to give them the main aim of the privacy policy without having to go through it.

1.3 Methodology

In order to accomplish what was mentioned above, research has been conducted in past attempts to make privacy policies more user friendly, this includes tools as well as legislation like the GDPR. Next, studies about the effects on users when changes were made to privacy policies have been looked at to see if users will welcome such changes or reject them. Moreover, studies on how users respond to different visualisations have been examined to gather data on what might be a good idea to use in this project. Finally, some research has been done on how to automate this process of labeling and categorising privacy policies in order to try something different or improve upon existing concepts.

The web application makes good use of all the information that has been gathered from the aforementioned research. It is important to learn from past successes and failures in order to avoid mistakes or mimic good techniques and practices to produce a good and functional result.

For the development of the classifier and the backend of the project, Python was chosen due to its ease of use and its vast number of libraries that will be useful for processing data, creating and training classifiers. Additionally, the existence of capable web frameworks in python that will help with the creation of the frontend and its connection to the backend are another great advantage. Next, Flask web framework was selected due to its relatively small footprint and its ability to create web applications with only the necessary requirements without including redundant libraries. Finally, for the frontend HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets), [Bootstrap](#) and Javascript were chosen in addition to some libraries for the creation of [pie charts](#) and [word clouds](#).

The first part of development was focused on the creation, training and testing of the classifier, which is the backbone of the entire project. The training and testing of the classifier has been done with the help of the [OPP-115](#) dataset [31] and it required a lot of trial and error until a set of suitable variables was found that produced usable results. The process of creating the classifier also involved the creation of the necessary preprocessing stages that processed the raw data that was given to the classifier. Next, the functions necessary to take advantage of the classifier were created. These functions connect with the frontend with the use of RESTful APIs that were created with the assistance of Flask. Finally, the frontend was made and connected to the backend using the RESTful APIs mentioned before. The frontend required a lot of polish before it was shown to the users as it will play a critical role in how the users perceive the web application and whether or not they will return to it in the future.

To conclude the project, a questionnaire was created and given to users to evaluate both their past experiences with privacy policies and their experience with the Privacy Policy Beautifier. This questionnaire plays a big role in the future of the project as it determines whether the project was successful at its task and what changes can be made to improve it further.

1.4 Chapter Outline

In the next chapter the concepts of personal data and privacy will be analysed, because they are the basis for the creation of this project, as well as the threats that loom over these concepts in our modern digital life.

In Chapter 3 a description of various other works will be given that are related to the aim of this project and will provide useful information to aid in the creation of the Privacy Policy Beautifier.

Chapter 4 will be focused on the web application itself. This will include all the steps that have been taken from the start of development until the final product is created. More emphasis will be given to the creation of the classifier and the web application, which are the two most important parts of this project.

In Chapter 5 a detailed description will be given on how the classifier and the application were evaluated. This evaluation will be crucial in determining whether the project will succeed in achieving its goal of making privacy policies more user friendly and making users more inclined in reading these privacy policies.

Finally, Chapter 6 will give a summary of the whole project along with the conclusion and findings of this project. Additionally, a discussion will be included analysing why this project was considered necessary and how it compares with other works that have been done before it. Moreover, some aspects of the project that can be improved in the foreseeable future will be mentioned along with how this project can help others in their work.

Chapter 2

Background

2.1 Personal Data and Privacy	6
2.2 Privacy Threats	9
2.3 General Data Protection Regulation (GDPR)	11

2.1 Personal Data and Privacy

As proclaimed by the United Nations General Assembly [28] and the Council of Europe [5] privacy is considered a fundamental human right.

The idea of privacy is very broad and multifaceted, making it very difficult and nearly impossible for theorists and academics alike to be able to determine or define it completely. Therefore, although referred to as a holistic idea, it is not defined as one specific construct, but instead through theorized and suggested so called ‘aspects’ such as:

- **Right to be let alone:** In the article “The Right to Privacy” [29] by jurists Samuel D. Warren and Louis Brandeis written in 1890, they discuss the idea of the “right to be let alone”. It is concluded that the “right to be let alone” refers to the right that a person has to be left alone or in isolation from others if they wish so, and the right to not be scrutinized or observed in a private setting like someone's own home.
- **Limited access:** Limited access concerns an individual's ability to be part of society without other people and/or organizations gathering information about their person [25]. Different theories exist, some of which think of privacy as a system that limits access to someone's personal information.

- **Control Over Information:** Having control over someone's own personal information is the idea that an individual, a group or an organisation have the right and the power to determine and decide on their own when, how and to what extent information concerning them is communicated to others. This aspect is ever more prevalent in the modern world of the 21st century, with the expansion of the internet and big data our control over information concerning our person is under threat. [4]
- **States of Privacy:** The four states of privacy are defined by Alan Westin [30] as:
 - **Solitude:** The physical disconnection from others.
 - **Intimacy:** Is the rather close and honest relationship of two or more people that originates from the isolation of a couple or group of people.
 - **Anonymity:** Is the yearning of individuals for times of "public privacy".
 - **Reserve:** Is the construction of a mental barrier to defend against undesired intrusion. This construction of a mental barrier needs others to acknowledge and respect a person's need or wish to limit the exchange of information that has to do with him/her self.

In regards to the mental barrier of reserve, Kristy Hughes [11] recognized 3 additional kinds of privacy barriers:

- **Physical barriers:** This includes physical objects like doors and walls that prevent other people from having access and experiencing the individual.
- **Behavioral barriers:** These are used to communicate with others either verbally through the use of words and language, or by non-verbal means, like body language, personal space or clothing that a person does not wish them to access or experience him or her.

- **Normative barriers:** These are usually social norms and laws that constrain other people from trying to access or experience a person.
- **Secrecy:** At times privacy is described as the option to have secrecy. When discussing privacy as secrecy it is usually seen to be a selective kind of secrecy where people choose to retain information to themselves (private), while they choose to make other information more public and not private. [25]
- **Personhood and autonomy:** Privacy can be considered as a necessary prerequisite for the expansion and conservation of personhood. Privacy may also be considered someone's ownership of both the physical and psychological aspects of his or her self. Additionally, privacy can be seen as a state that enables autonomy which is closely related to that of personhood. [16]
- **Self-identity and personal growth:** Privacy can be viewed as a precondition for the expansion of a sense of self-identity. To be more specific, privacy barriers are instrumental in this process. The control of these barriers enables one to define the boundaries of the self, which in turn helps define the self. This essentially means being able to control and regulate contact with others. Additionally, privacy may be considered the state that hosts personal growth, a process necessary to the expansion of self-identity. [16]
- **Intimacy:** Personhood theory describes privacy as being a necessary part of the way that people have reinforced relationships with other people. This may be the case because an integral part of human relationships involves people willingly disclosing most, if not all, of their personal information.[25]
- **Personal / Physical Privacy:** This could be described as stopping invasions into someone's physical space or their place of residence. [12]
- **Organizational:** Corporations, government agencies, societies, groups and other organisations frequently wish to keep their practises, intentions or secrets from becoming known to other organisations or people. In order

to achieve this, many security practices and protocols are placed in hopes of keeping private information confidential.

Privacy is fundamental to who we are and how we grow and define our “self”. In the era of the internet and big data the importance of privacy has come to the foreground of everyone's mind. The free flow of information on the web has made it extremely hard to keep track of one's personal information. The aforementioned mental and physical barriers defining “Personhood and autonomy”, no longer apply when it comes to technology.

Being able to decide when, where, how and more importantly with whom to disclose any personal information is vital in the construction of the “self” and in building and maintaining strong relationships with other humans.

2.2 Privacy Threatened

The modern world is constantly evolving technologically, making information, instant access and entertainment through the web thoughts at people's top priorities. Despite all the obvious benefits of this new era of technology, there are many hidden threats that are lurking in the dark. It is easy to ignore all the warnings as the majority of the world now has adopted a new way of life centered around technology via personal computers, smart phones, social media and all sorts of programs, applications and other conveniences.

All of these applications and systems that people use on a daily basis store information about their users, and in many cases a lot more information than users realise. Even if people are aware what information these organisations store about them it is not always clear what that information is used for or who has access to it, either intentionally or not.

There have been many major data breaches in big corporations in recent years, for example:

- Yahoo breach in 2013-2014 (disclosed in 2016-2017), an attack compromised more than 3 billion user accounts, usernames, emails, telephone numbers, birth dates, encrypted passwords and some security questions and answers were taken.
- FriendFinder Networks breach in 2016, affected 412,000,000 accounts with usernames, passwords, email addresses, and some other details discovered. This also included personal information from deleted accounts.
- Marriott-Starwood breach, hackers had access to the Starwood database for four years (2014-2018) before the breach was discovered. During this time, passports, telephone numbers, emails, and some credit card details were being stolen.
- Myspace breach in 2016 resulted in 360 million accounts being compromised and email addresses and passwords were stolen and posted in a hacker forum. This particular incident shows that information remains available and vulnerable even years after it was given to an organisation (Myspace was popular in the late 2000s).
- Under Armour breach in 2018 had 150 million accounts affected from the MyFitnessPal app. The information that were stolen included usernames, passwords and email addresses.
- Equifax breach occurred in 2017 when hackers exploited a web server's vulnerability. This was reportedly preventable but the company's security practises were not efficient and its systems were quite old. In this incident the information stolen contained names, birth dates, social security numbers, addresses and in some cases credit card numbers.

These major breaches all happened between 2013 and 2018, a span of 5 years, but they do not tell the whole story as many more breaches happen very frequently but on a smaller scale. This issue is even more significant when we consider the increasing volume of information each individual is giving to these organisations that are obviously vulnerable to attacks and leaks. In addition, a

large number of users aren't aware of the information they are providing to these organisations [22] or what this information is used for and how at risk it is.

2.3 General Data Protection Regulation (GDPR)

Realising the severity of this issue and its increasing importance and impact on the modern age of technology, the European Union decided to step in and attempt to protect all European citizens by voting and applying the EU General Data Protection Regulation (GDPR) [23] that was proposed on the 25th of January 2012, before being entered into force on the 24th of May 2016 and was finally implemented on the 25th of May 2018. After being implemented, its provisions were directly applicable in all member states.

The General Data Protection Regulation replaces the older Data Protection Directive that was enacted in October 1995. Its main goal is to give individuals control over their personal data and to make the regulatory environment simpler for international business by unifying the regulation within the European Union. The GDPR [23] is made up of 99 articles and 173 recitals, some of them are Lawfulness of Processing (Article 6), Conditions for consent (Article 7), Right to rectification (Article 16), Right to Erasure (Article 17), Right to Object (Article 21), Right of Access (Recital 63), Restriction of Processing (Recital 67) and much more.

The introduction of the GDPR has caused some issues for companies and organisations as they were now forced to make the necessary changes to their policies, practises and generally their way of operation as to be GDPR compliant. This was of the utmost importance because failing to comply with this new regulation and violating the constraints placed by it would result in severe consequences for the organisation, such as legal and financial penalties. These penalties can reach amounts of up to 20 million euros or 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher, as mentioned in article 83.

Some noticeable examples of these enormous fines are shown in table 2.1 [8]:

Organisation	Fine amount (€)	Date	Quoted Article
Marriott International Inc	20.450.000	30/10/2020	-Art. 32 GDPR
British Airways	22.046.000	16/10/2020	-Art. 5 (1) f) GDPR -Art. 32 GDPR
H&M Hennes & Mauritz Online Shop A.B. & Co. KG	35.258.708	30/10/2020	-Art. 32 GDPR
Wind Tre S.p.A	16.700.000	13/07/2020	-Art. 5 GDPR -Art. 6 GDPR -Art. 12 GDPR -Art. 24 GDPR -Art. 25 GDPR
TIM (telecommunications operator)	27.800.000	15/01/2020	-Art. 5 GDPR -Art. 6 GDPR -Art. 17 GDPR -Art. 21 GDPR -Art. 32 GDPR
Austrian Post	18.000.000	23/10/2019	-Art. 5 (1) a) GDPR -Art. 6 GDPR

Table 2.1 Major fines given for non-compliance with the GDPR

Despite its importance and noble goal, the GDPR has proven to be a great challenge when it comes to small and medium-sized companies who claim that the costs to be GDPR compliant place a heavy burden on them. Additionally, it is sometimes unclear how these regulations and rules apply to new and emerging technologies leading to confusion [6].

Chapter 3

Related Work

3.1 Introduction	15
3.2 Research on the topic of privacy policies	15
3.2.1 Disclosing Personal Data Socially - An Empirical Study on Facebook Users' Privacy Awareness	16
3.3 Research on the topic of privacy policies and GDPR	18
3.3.1 The Privacy Policy Landscape After the GDPR	19
3.3.2 The Case for a GDPR-specific Annotated Dataset of Privacy Policies	20
3.3.3 Evaluating privacy policies on web platforms based on the GDPR	22
3.4 Research on the topic of improving the visualization of privacy policies	22
3.4.1 Effects on privacy policy visualization on users' information privacy awareness level (instagram)	24
3.4.2 Polisis Automated Analysis and Presentation of Privacy Policies Using Deep Learning	26
3.4.3 PrivacyCheck Automatic Summarization of Privacy Policies Using Data Mining	27
3.4.4 Quantifying the Effect of In-Domain Distributed Word Representations	29

3.4.5 Towards usable privacy policy display and management	30
3.4.6 Unsupervised Topic Extraction from Privacy Policies	32

3.1 Introduction

In this chapter, the main research papers used and studied to make this research will be summarised. These research papers were mainly focused on three things, privacy policies, the General Data Protection Regulation (GDPR) and whether altering the appearance or structure of a privacy policy makes it more understandable, easy to read, and generally more user friendly as well as how this affects the end users. More specifically, the effects of the introduction of the GDPR will be seen, in addition to some attempts to evaluate privacy policies based on it. Finally, a comparison will be made on the different approaches made to improve upon existing privacy policies whether that is with automated or manual methods.

3.2 Existing works on privacy policies

Privacy policies have been getting increasingly large and hard to read in recent times. This change probably happened for a number of reasons, some of which might be the number of laws and regulations written from various governments or organisations to force companies and businesses to be more honest for the protection of the consumer. This in turn made privacy policies more complex, time consuming and generally uninviting for the average user. It is obvious from everyday life that most people have probably never read a privacy policy before which is concerning to say the least.

3.2.1 Disclosing Personal Data Socially - An Empirical Study on Facebook Users' Privacy Awareness

The rapid increase of users in social networks services like Facebook has ignited concerns over privacy due to the enormous amounts of personal data these services collect from their users. This paper focuses on those users by conducting a survey on 210 Facebook users. The aforementioned survey reveals that a majority of active users on Facebook share a large amount of personal information. Additionally, these users are usually not aware how visible their information is to strangers. Moreover, the privacy policy and terms of use that all users must accept or agree with to use the service are largely unknown or not understood.

The privacy concerns arise when users reveal identifiable data about their person to users or people that they would not trust in the real world. Studies have shown that some students are aware of the risks of disclosing such information online and know of ways to limit the visualization of this information but have taken no steps to do so. Other studies have seen that users are not aware of these concerns or they believe that their personal risk is not significant enough.

The study showed that the majority of users shared a large amount of personal information as seen in Table 3.2.1.1 and that a lot of them had chosen to make that information available to their friends (63%) but still a large number of users (34%) made that information available to all users part of the same platform.

Users seemed to be slightly worried about their privacy when it came to using the internet and they were aware of possible threats like identity theft or having their credit card number stolen. Despite those worries, users still displayed a level of trust when it came to the other internet users. Additionally, the participants seemed to be aware and accustomed with the notions of data protection and security. It was also observed that almost all users (94%) were aware they could change their privacy settings and most of them (84%) claim to have done it.

Questionnaire Item	n	%
Real name	208	99
Profile picture	206	98
Birthday	186	89
Home town	186	89
E-mail address	174	83
Education information	169	80
Photos of one's self	158	75
Photos of one's friends	130	62
Relationship status	124	59
Sexual orientation ("interested in")	103	49
Favorite music, movies, etc	70	33
Contact phone number	69	33
Activities / interests	67	32
Partner's name	55	26
Street address	38	18
Website	25	12
Political views	20	10

Table 3.2.1.1 Pitkänen and Tuunainen [22]

In conclusion, the privacy policy of any website, especially that of social network services is very important as it contains information about how the users' personal data is being processed and utilised, but also because it is a consent form that users agree to. Additionally, reading the privacy policy doesn't always increase the users' privacy awareness as privacy policies tend to be very long and hard to understand by the average user.

3.3 Existing works on privacy policies and GDPR

With privacy becoming more threatened and with applications and sites requesting or demanding an increasing volume of personal data from their users, the European Union decided that users needed to be protected and informed about what information they are providing, how it is being processed and who has access to it. All this came in the form of the General Data Protection Regulation which aimed to protect users data. As stated in the previous chapter the sole purpose of the GDPR is to protect the users and give them more options and control over their personal information. The GDPR is enforced by the EU and if companies or organizations that fail to follow the rules and regulations are punished with heavy fines.

In this section some research papers will be mentioned and analysed that have to do with the GDPR itself. Firstly, a paper that has to do with the effect that GDPR has had on privacy policies both inside and outside the European Union [18]. This paper is very important as it helps to better understand whether or not the introduction of the GDPR has actually made any impact in the world of privacy policies as well as if this effect is contained inside the European Union or if it is affecting companies and organisations outside the EU. The knowledge of whether the GDPR has forced companies or organisations not in the EU to adopt the standards and adjust their privacy policies is very significant for two main reasons. Firstly, because in our day and age of technology and globalisation european citizens whom the GDPR is trying to protect, are using and interacting with websites and applications from organisations that may not be based inside the EU. Secondly, if the enforcement of the GDPR is applying pressure to organisations worldwide to be more user friendly and are forced in a way to use more honest business models and practises, that means that all users around the world benefit.

Afterwards, an analysis is carried out on why GDPR-specific annotated datasets of privacy policies should exist, and how effective are the ones that we already have [7]. Having large datasets of something specific is always useful when it

comes to automation, as it allows for the creation of data-driven algorithms and the training of neural networks through large amounts of training data for better and more accurate results. Finally, a tool made to evaluate privacy policies based on the GDPR [1] will be summarised as it is useful for users and organisations to be able to see in an easy to read and understand way if a privacy policy fares well when it comes to the many rules and regulations of the GDPR.

3.3.1 The Privacy Policy Landscape After the GDPR

This paper focuses on the impact the introduction of the GDPR has had on the world of privacy policies. By creating a corpus of 6.278 unique English-language privacy policies from both inside and outside the European Union and comparing their versions from before and after the enforcement of the GDPR. After gathering and analysing the test results, a conclusion was reached that the introduction of the GDPR has been the reason for a major change in the privacy policy landscape with most changes being in EU-based websites. Furthermore, it was observed that privacy policies had become significantly longer in length, probably to cover and satisfy the new regulations. Despite being more extensive, the new privacy policies had 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 becoming more extensive and descriptive. But that always came at the cost of readability and clarity of the privacy policy.

More specifically, the privacy policies were tested to see changes in five dimensions: presentation, textual features, coverage, compliance, and specificity. The changes were more significant in EU-based websites which were the GDPRs primary target but various changes and improvements were noticed at a global scale showing the effect and reach of the GDPR. Finally, even though the privacy policies got longer after the introduction of the GDPR (a fifth longer in the Global set and a third longer in the EU set), the presentation, and readability was

improved as it can be seen by the positive trend in user experience for the EU policies.

3.3.2 The Case for a GDPR-specific Annotated Dataset of Privacy Policies

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. Their paper revises existing and related datasets to try and give an analysis of how they could be modified or changed in order to make it possible for them to be used in various different machine learning techniques. This will have a considerable impact in training models and systems that will be able to test the compliance of different privacy policies automatically.

The paper mentions that one suggestion to help companies be compliant with various regulations was to stop using natural language to express privacy policies and instead use some formal language [13]. This has the appeal of being processed much easier by machines to be transformed directly into a structured database. Despite that potential, it has not seen much adoption in the market. On the other hand, various natural language processing techniques have been introduced that try to solve this issue. The paper says that in the beginning, these techniques used unsupervised learning due to the lack of annotated datasets. Afterwards, when the dataset OPP-115 [31] was created, a chance for (semi)-automatic processing was created. This opportunity cleared the way for applications like Pri-bot [10] to make an appearance.

The paper highlights that with the introduction of the GDPR, natural language processing techniques can be very beneficial especially to small businesses and enterprises that are trying to be compliant. But these techniques are very reliant on annotated datasets for training. Before applying any of these techniques the paper suggests that these four considerations that risk introducing some major bias should be taken into account:

1. **Impact of new elements:** The current datasets do not address many of the GDPRs very specific elements. These elements are

very important when it comes to compliance with the GDPR and the fact that they are not covered fully or not at all by current datasets might cause issues.

2. **Impact of multi-linguality:** There are 24 official languages within the European Union. This is an issue when considering smaller companies that may provide services in their local language. This introduces the need for a more strict dataset available in at least two languages.
3. **Impact of domain shift due to the type of companies:** The GDPR is a law that affects any business whereas the current data sets only focus on privacy policies from popular websites, which means smaller companies and organisations are excluded.
4. **Impact of domain shift due to adaptation to the GDPR:** After the introduction of the GDPR, large modifications took place in a large amount of privacy policies. This was noticeable by the large amount of emails from services and sites informing users of changes made to the companies' privacy policy. Machine learning tools are very sensitive to such changes and that means that the ones trained before the introduction of the GDPR should at least be measured to see whether they have been affected or not.

The paper concludes that the current datasets should be revised and the four considerations mentioned above should be taken into consideration before applying any algorithms trained on them. Moreover, these datasets can be updated with the new GDPR elements that they are missing. Having said that, the impact on policies in different languages as well as companies whose policies were already annotated and knowing that those policies underwent major changes after the introduction of the GDPR should be analysed. If the analysed results show a significant impact, then a new GDPR-specific dataset should be considered.

3.3.3 Evaluating privacy policies on web platforms based on the GDPR

This particular paper focuses on creating a tool that can be used by various businesses and organisations to be able to check that the privacy policies they create are compliant with the GDPR. Each organisation is responsible to make sure that their privacy policy doesn't violate any of these newly introduced regulations as they risk getting heavy fines. This process is usually very costly especially for small to medium sized organisations, putting them at risk of getting fined. With a tool like this, smaller organisations can have a better idea of what needs to be done for their privacy policy to be GDPR compliant.

The paper starts by conducting some research on the GDPR itself to determine key words and phrases that privacy policies need to include in order to be GDPR compliant. Afterwards, a crawler was implemented to find the url/page in which the privacy policy of a site is located. After finding the privacy policy, a parser is used to extract the required text from the webpage which is then analysed and processed in combination with the list of key words and phrases that was constructed before. Finally, a score is given to the privacy policy determined by how well it covers all necessary points of the GDPR, as well as giving an option for a more detailed analysis that shows which points have been included and what is missing from the privacy policy.

Tools like these are very useful and can be beneficial for both organisations and end users. With these tools finding issues, red flags or unnoticed details becomes a lot easier, so organisations can adjust and modify their privacy policies and their practises accordingly, and end users can know whether they should avoid or take caution when using certain websites or applications.

3.4 Research on the topic of improving the visualization of privacy policies

This section focuses on papers and research that has been done concerning the visualisation and presentation of privacy policies. With regulations and laws putting more pressure on companies and organisations, now more than ever, to

be more transparent and clear about their processes, privacy policies have become longer and more complex than ever. As mentioned in 3.3.1, the introduction of the GDPR had a similar effect but in contrast to other regulations, a more positive user experience was observed in updated privacy policies trying to comply with the GDPR. Despite this positive step forward, there is still a long way to go before the average user is compelled to read a privacy policy. It is well known that very few users spend the time and effort to read a privacy policy despite agreeing to everything with a single click. This happens due to their length, complexity and vagueness [17, 19].

There have been many attempts to try and improve upon the visualisation of privacy policies, some manual and some automatic. Firstly, a study proposing 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 [26]. Other approaches use more automated means to analyze, process and transform the privacy policy, in ways that will be easier for the average user to understand. These methods may use deep learning and graphs for representing the necessary information [10] or they might use 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 [32]. 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 [15], or to extract topics from privacy policies and even unveil new ones that supervised methods might have missed [24]. Finally, there is also a paper that proposes to give the users the power to make calculated choices on the distribution of their personal information [3]. These different techniques and methodologies will be analysed further below.

3.4.1 Effects on privacy policy visualization on users' information privacy awareness level (instagram)

In this paper the idea is that different visualizations of the same privacy policy may lead to different levels of privacy awareness for the end users. To examine this hypothesis, the privacy policy of instagram was used. The privacy policy was represented in three different ways:

1. **Unchanged:** The original privacy policy in the conventional textual representation.
2. **Tag clouds (WordBridge):** Tag clouds is a widely used technique, used to help users consume the contents of a document. In this instance the WordBridge (Figure. 3.4.1.1) technique was used, which belongs in the Tag Clouds family. WordBridge uses nodes and links that are both tag clouds, to represent the entities of the document (nodes) and the relationships between them (links) [14].

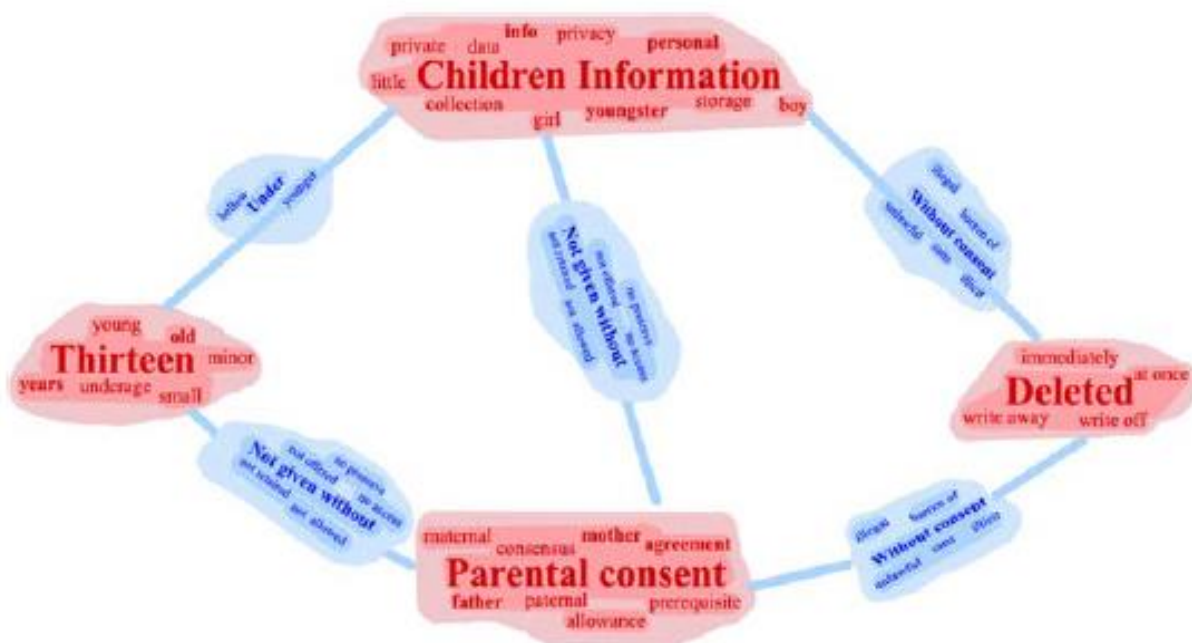


Figure 3.4.1.1 Tag Clouds representation of privacy policy information

3. **Document Cards:** Document Cards (Figure. 3.4.1.2) uses a combination of important keywords/terms and images to represent the content of a document in a more compact overview [27]. The images draw the interest of the user and the keywords and terms provide further explanation.

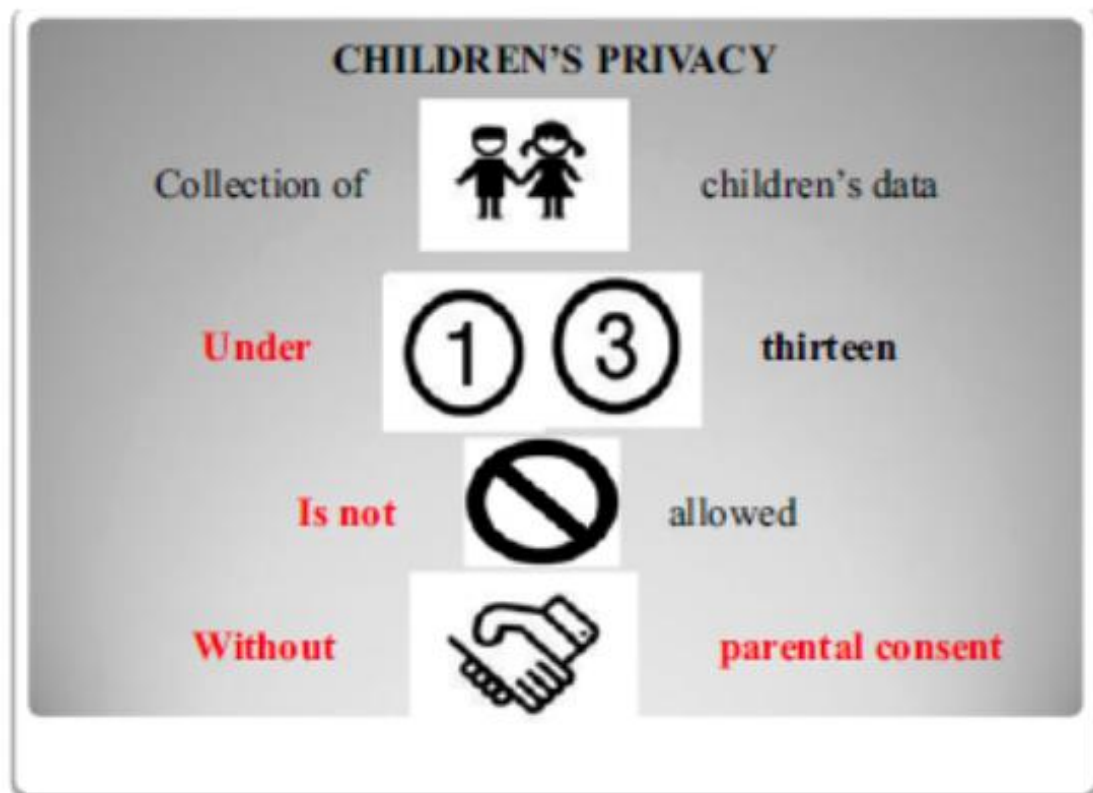


Figure 3.4.1.2 Document Cards representation of privacy policy information

This research involved both technically experienced students and students from other non-technical fields of study. These students were given a questionnaire to determine their privacy awareness. Afterwards, the students were divided in three groups and were asked to read the privacy policy. One group was given the original conventional textual privacy policy, the second group was given the WordBridge version of the privacy policy and the final group was given the Document Cards version of the privacy policy. Upon reading the privacy policy the students were then asked to answer another questionnaire to determine any changes to their privacy awareness.

The paper concludes that after reading the privacy policy, users had a higher privacy awareness, this is due to the fact that privacy policies contain all the necessary and required information that they should, but users rarely read them as shown in a variety of other research papers. Additionally, it was apparent that the two new visualization techniques resulted in a higher policy awareness than the traditional textual representation. Finally, from the two new techniques used, the most effective in raising privacy awareness was the “Privacy Policy Cards” that made use of the Documents Cards.

3.4.2 Polisis Automated Analysis and Presentation of Privacy Policies Using Deep Learning

This paper shows the attempt of researchers to overcome the hurdle of scalability when it comes to creating small notices based on information that has been derived from privacy policies. In order to tackle this lack of usable and scalable tools, pribot was created to automate the process so companies, users, regulators and researchers can save time and effort.

Pribot is an automated framework built for analysing privacy policies. It was created with the help of 130K privacy-centric language model and a novel hierarchy of neural network classifiers that analyze both high level and fine grained details from privacy policies. Additionally, Pribot is capable of answering both structured and free form queries. The structured querying assigns privacy policy icons from privacy policies automatically with an accuracy of 88.4%, whereas the free form question answering application of PriBot is capable of providing a correct response in its top three results for 82% of the test questions.

Other attempts have been made to create UIs that will present privacy policies in a more user friendly way. These attempts include machine readable formats, nutrition labels, privacy icons and short notices. Despite these efforts, they all face the same obstacle, the amount of time and effort of adding new notices to existing policies. PriBot is capable of breaking up a privacy policy into smaller

self-contained segments of text. It then, automatically annotates these segments with a set of labels that describe its data practices, this is done with high accuracy as shown by the scores PriBot achieved in its tests (Table. 3.4.2.1).

Label	Prec.	Recall	F1	Top-1 Prec	support
Data Retention	0.83	0.66	0.71	0.68	88
Data Security	0.88	0.83	0.85	0.79	201
Do Not Track	0.94	0.97	0.95	0.88	16
1st Party Collection	0.79	0.79	0.79	0.79	1211
Specific Audiences	0.96	0.94	0.93	0.93	156
Introductory/Generic	0.81	0.66	0.75	0.75	369
Policy Change	0.95	0.84	0.93	0.93	112
Non-covered Practice	0.76	0.67	0.6	0.6	280
Privacy Contact Info	0.9	0.85	0.88	0.88	137
3rd Party Sharing	0.79	0.8	0.82	0.82	908
Access, Edit, Delete	0.89	0.75	0.87	0.87	133
User Choice/Control	0.74	0.74	0.69	0.69	433
Average	0.85	0.79	0.81	0.8	

Table 3.4.2.1 Classification results at the category level for the Segment Classifier

3.4.3 PrivacyCheck Automatic Summarization of Privacy Policies Using Data Mining

This paper is about a tool made to create summaries of privacy policies automatically. The tool is called PrivacyCheck and it is used as a Chrome browser extension to summarize any HTML page that contains a privacy policy. The need for a tool like this emerged due to the fact, also shown from many previous papers, that most users don't read the privacy policies provided by the different websites. Despite users being aware of privacy policies and their importance, most people don't spend the time and effort required to read them. This is due to

various reasons, most importantly the amount of time it would take to read all of the privacy policies that a user would encounter in their daily lives, the length of the privacy policy and the fact that most privacy policies are too hard to understand and comprehend for the average user.

PrivacyCheck is an add-on browser that requires the user to insert the URL of the page where the privacy policy is located. Afterwards, it presents to the user a quick and easy to understand summary of the contents of the privacy policy. This novel representation of contents is done with the use of appropriate icons and colors to indicate the topic and the level of risk concerning that aspect of the privacy policy. The summary is done based on a list of 10 privacy factors that are represented by the following questions:

1. How does the site handle your email address?
2. How does the site handle your credit card number and home address?
3. How does the site handle your Social Security number?
4. Does the site use or share your personally identifiable information for marketing purposes?
5. Does the site track or share your location?
6. Does the site collect personally identifiable information from children under 13?
7. Does the site share your information with law enforcement?
8. Does the site notify you or allow you to opt out when their privacy policy changes?
9. Does the site allow you to edit or delete your information from its records?
10. Does the site collect or share aggregated data related to your identity or behavior?

PrivacyCheck then automatically makes a prediction for each privacy factor's risk value mentioned above, using a classification data mining model (supervised machine learning). These risk values are then presented to the user as seen in Figure.3.4.3.1.



Figure 3.4.3.1 A snapshot of the Privacy Check Chrome extension

3.4.4 Quantifying the Effect of In-Domain Distributed Word Representations: A Study of Privacy Policies

This paper evaluates the advantages of having privacy specific word embeddings when using Natural Language Processing to extract or summarize statements from privacy policies. To accomplish this evaluation, a corpus of 150 thousand privacy policies is used to build word vectors using unsupervised techniques. These privacy specific embeddings are created with the hope of accelerating and helping future research.

The paper contributes in various fields. Firstly, an investigation concerning the utility of in-domain word embeddings found that they help over generic word embeddings and have improved performance in segment-labeling in the privacy policy. Secondly, an investigation is conducted on the relationship between

dimensionality of the word embeddings and segment labeling performance. Thirdly, another investigation is conducted to determine how many privacy policies are needed to train expressive word embeddings, which in this case appeared to be around 20,000, as the F1 score plateaus when given more privacy policies.

3.4.5 Towards usable privacy policy display and management

The goal of this paper is to indicate the approach within the PrimeLife project for designing user friendly privacy policy interfaces for the PrimeLife Policy Language (PPL) and explain the information that can be learned when designing and implementing interfaces for displaying and managing privacy policies. The powerful features given by PPL had to be handled in some way, so the browser extension “Send Data?” was designed and developed specifically for this purpose. These new features allow users to make calculated choices on the distribution of their personal data, as well as the new feature of “on the fly” privacy management. The new features also include preconceived levels of privacy settings and simplified selection of anonymous credentials.

In the real world, away from the computer or smartphone, people regulate their privacy almost completely automatically. They decide where, when, why, how and to whom they share their personal information. This can be seen in everyday interactions, for example people will share different information about themselves with their coworkers in comparison with their family members, partner or close friends. This process is something that happens automatically and every person does this almost everyday without having to spend time and effort. The difficulty here is to convert this process of automatic choice, management and understanding of personal privacy to the digital world.

The “Send Data?” browser extension prototype (Figure. 3.4.5.1) presents to the user fundamental elements of a service provider’s privacy policy in an easy to understand and user-friendly way. Furthermore, it shows to the user how their

own predefined privacy preferences match service provider's privacy policy when their personal data is being requested.

Send Data?

Your data will be sent and used for the following purposes

Data attributes	Purposes					Conditions
	Administration	Contact	Feedback	Marketing	Payment	
Name - Certified By: Driver's License [Swedish] - ... Inga Vainstein	> Ex	>>	> Ex	> Ex >>		
Credit Card - Certified By: Visa Credit Card [My private...] 1234 5678 9012 3456 Exp: 2012-01-12					> V	10 days
E-Mail: 	> Ex >>	>>	> Ex	> Ex >>		10 days

> Data will be sent to:
 Ex Example.com ([Privacy Policy](#)) (store.example.com, contact@example.com)
 V Visa ([Privacy Policy](#)) (www.visa.com, customersupport@visa.com)

>> Data will be forwarded to others
 Does not match your privacy settings
 Matches your privacy settings

Privacy policy matching results

[Click to see all mismatches](#)

My current privacy settings:
 Medium Privacy Settings
☐ Accept mismatch
 for this transaction only

Cancel Send

Figure 3.4.5.1 An alternative redesign for the seventh iteration cycle of the “Send Data?” prototype

In conclusion, it is clear that the powerful features provided by the PPL add complexity and decrease usability. Additionally, developing a user-friendly interface for such a language is no easy task. User testing has shown that users understand the fundamental aspects of the “Send Data?” prototype. Moreover, it has been observed that the use of colour in a 2D table is an adequate way of illustrating mismatches between the users’ privacy preferences and the service provider’s privacy policy. Furthermore, the novel feature of “on the fly” privacy

settings appeared to be understood and welcomed by the users. Despite these positive results, there are many improvements to be made and hurdles to overcome in future iterations. These include displaying user friendly privacy policies in smartphone devices with smaller screens, as well as getting informed consent from users in an age of ubiquitous computing where interconnected devices will be practically invisible and interaction might occur without the use of visual interfaces.

3.4.6 Unsupervised Topic Extraction from Privacy Policies

This paper focuses on the use of unsupervised learning techniques for automatic topic modeling for large scale corpora. The paper suggests unsupervised learning techniques as it has certain advantages that help it stand out. Some of the main advantages, incorporate but are not limited to the ability to analyze any new corpus with significantly less effort than it would require a supervised learning technique, the ability to observe modifications in topics of interest as times goes on, and the ability to recognize finer grained topics in these privacy policies. The use of an unsupervised learning technique, like topic modeling used here, also negates the need for specific tagged datasets that take a lot of time, effort, and money to create as to be able to train supervised learning models.

Unsupervised learning requires strict validation to make sure that the list of extracted topics is not just a subset of the actual list. The topics extracted have been compared and validated based on other papers and work done in the past. The 36 extracted topics have been manually mapped onto the reference list, which shows a big overlap between the topics which encourages the idea that the unsupervised model is capable of producing results similar to those produced by equivalent supervised techniques. Additionally, some of the topics extracted had not been seen in previous works which could possibly suggest that the unsupervised technique was capable of finding new topics of interest that have been missed or overlooked by supervised learning techniques. These newly

found topics of interest may be the result of the recently introduced General Data Protection Regulation (GDPR) [23]. Finally, unsupervised learning may not need human effort for annotating the data like it is needed for supervised learning, but it does require some manual post processing by experts to make sure that all topics extracted and summaries produced for the different topics are valid and applicable. It is important to note that this manual task is immensely smaller than what is needed for the annotation of data for the supervised learning techniques.

Chapter 4

Privacy Policy Beautifier and how it works

4.1 Introduction	34
4.2 Training The Classifier	36
4.2.1 Background	37
4.2.2 Training	37
4.2.3 Previous Attempts	43
4.2.4 Final Classifier	44
4.3 Privacy Policy Beautifier	45

4.1 Introduction

As part of this thesis the web application Privacy Policy Beautifier was designed and implemented. The purpose of this web application is to provide an easy way for end users to make long and complicated privacy policies easier to read and understandable in an automated way. Privacy Policy Beautifier utilizes supervised training techniques to create a model whose purpose is to classify the different parts of a privacy policy. These classified segments are then presented to the user with different colors to make it easier for the user to find the information they are most interested in. Additionally, Privacy Policy Beautifier presents information concerning the privacy policy in other ways as well, like pie charts and wordclouds, to help summarize key points of the privacy policy. Finally, the

web application lists the appearance of key terms from the General Data Protection Regulation (GDPR) [23] inside the privacy policy.

This web application was motivated by the lack of privacy awareness that has been observed in end users when it comes to using services from the internet [17, 19, 22]. Despite the efforts of many governing bodies with laws and regulations [5, 23] written and implemented specifically for this reason, the problem has seen some improvement but it still endured [18].

In our day and age it is vital that people know what information they share online, how that personal information is stored, processed, used, or even sold to others, like advertising companies. As these online services play an ever increasing role in people's daily lives, it is up to every person individually to make sure they protect themselves and their personal information, as it is not always easy for governments or agencies to protect them [20]. But even though most people are aware of the risks and dangers of their personal data being available to the public or used by companies, they are still not taking action to change this situation [22].

The problems mentioned above are the reasons why tools, like Privacy Policy Beautifier, are needed. The aim is to find alternative ways to present a privacy policy to the end users as to encourage them to read it, or at least the parts that they find more important to them. 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 [26]. This is why the Privacy Policy Beautifier makes use of colors, tables and pie charts as can be seen later on.

The tools that are created to illustrate privacy policies differently must be able to accomplish this task automatically. This is because it would be impossible to do it manually, considering the enormous number of privacy policies that exist, the number of new privacy policies created and the fact that privacy policies constantly change depending on the companys' intentions, or even by the

introduction of new laws or regulations. Automating this process is no easy task but natural language processing and machine learning has made it possible. As seen from previous attempts, both supervised and unsupervised learning can be effective under the right conditions [10, 24, 32]. Privacy Policy Beautifier uses supervised training, which will be analysed further later on.

The architecture used in this project is a client server model (Figure 4.1.1). The client, or in this case the user, sends the URL or the text of a privacy policy he/she is interested in to the server. The apache server then processes that information and gives it to the pre-trained model for classification. The apache server and the model are located on the same VM. Next, the response given by the model is organized and sent back to the client. Finally, the browser at the client's side processes this response and presents it to the user in a visually pleasant and easy to understand way.

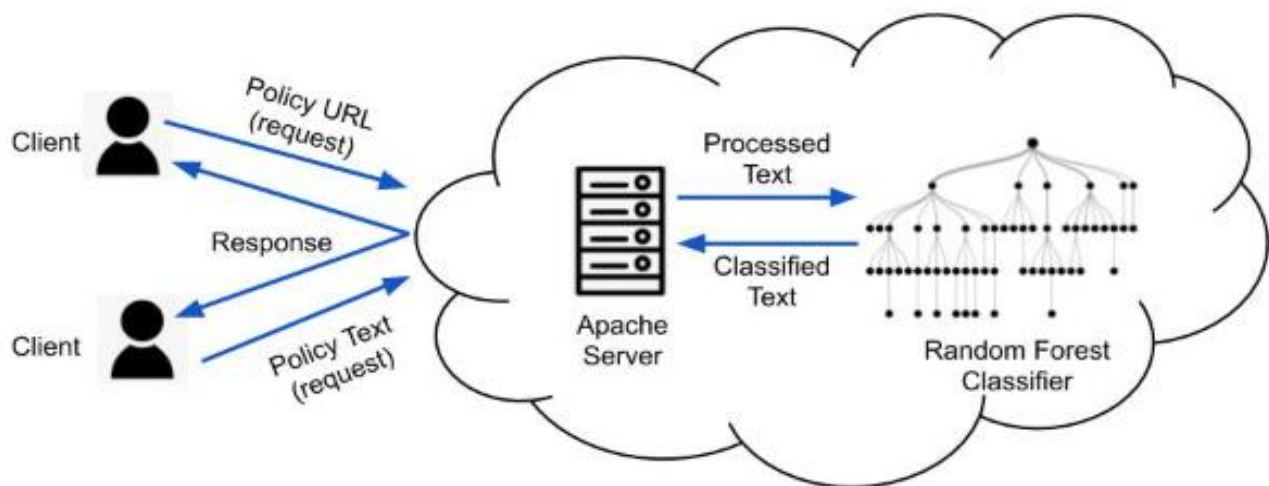


Figure 4.1.1 System Architecture

4.2 Training The Classifier

In this subchapter an explanation will be given on why a classifier was needed to tackle this problem, as well as how and why many decisions were made concerning the different parameters used by the classifier. Additionally, a description will be given about the training process of the classifier with the different technologies, algorithms and techniques used along the way. Finally,

some early attempts will be mentioned, that were made at the start of the project before being replaced by the final classifier that was used for the latest version of the Privacy Policy Beautifier.

4.2.1 Background

To help end users find what they are looking for, the privacy policy is divided into 10 different classes. These classes have been taken from the classification of privacy policies analysed in the OPP-115 dataset [31]. The 10 different classes are as follows: “Other”, “Third Party Sharing/Collection”, “User Choice/Control”, “Policy Change”, “Data Security”, “First Party Collection/Use”, “User Access, Edit and Deletion”, “Data Retention”, “Do not Track”, “International and Specific Audiences”. These categories are used to group the different information contained within the privacy policy. The dataset contains 115 privacy policies, as indicated by its name, which is not a large number considering the amount of privacy policies that exist at the moment. In addition, these policies were manually processed and divided into segments. Then, the segments were classified using the previously mentioned classes. For the Privacy Policy Beautifier to be effective, the process of dividing a privacy policy and classifying the segments needed to be automated. To accomplish this task, a classifier was created and trained using the OPP-115 dataset. This classifier can be given a sentence from any privacy policy and will attempt to position that sentence into one of the 10 categories as accurately as it can.

4.2.2 Training

As seen in Figure 4.2.2.1, to train the model the necessary data was taken from the OPP-115 dataset and moved to the pre-processing stage. Here, the data was put through a process of removing unnecessary information that would not be beneficial to the model. The first things to be removed were the stop words from the English language. Stop words are words that provide no useful information or context concerning the text that we want to process. The following words can be considered

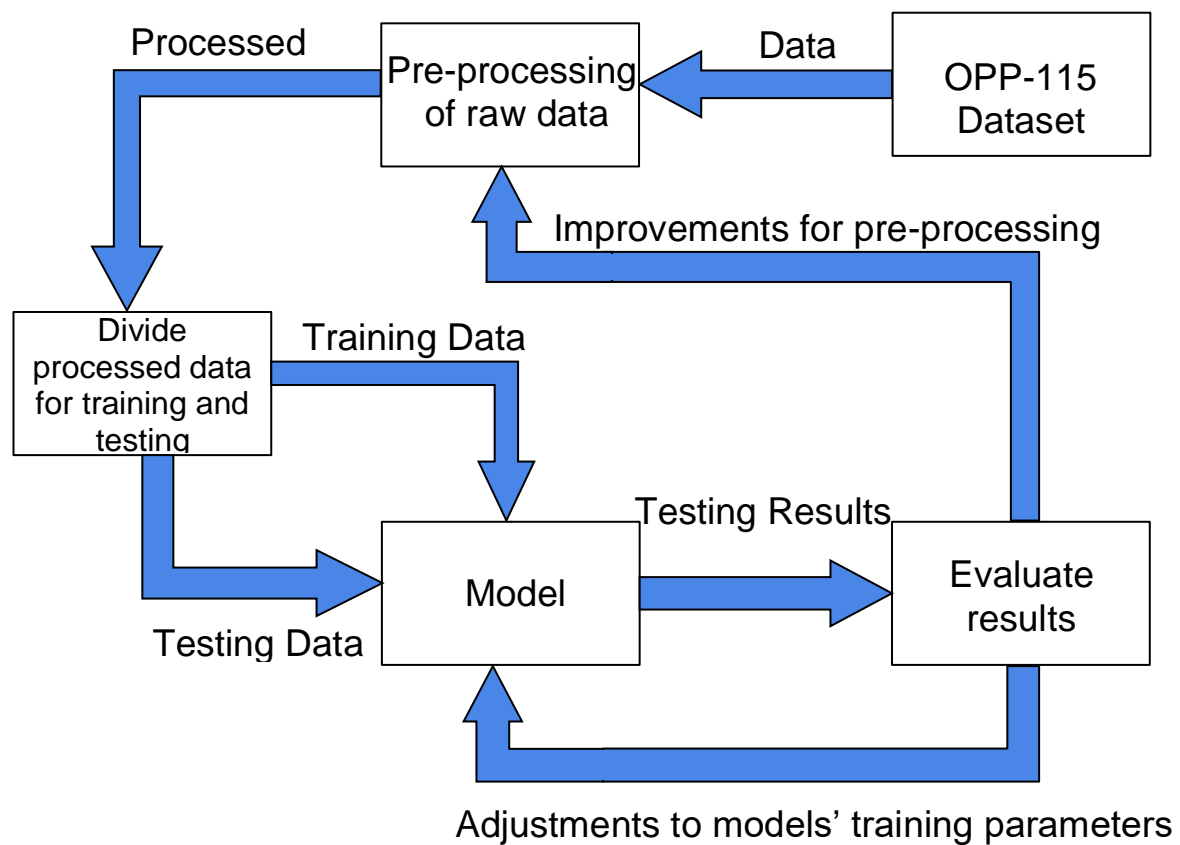


Fig. 4.2.2.1 Training and Evaluation

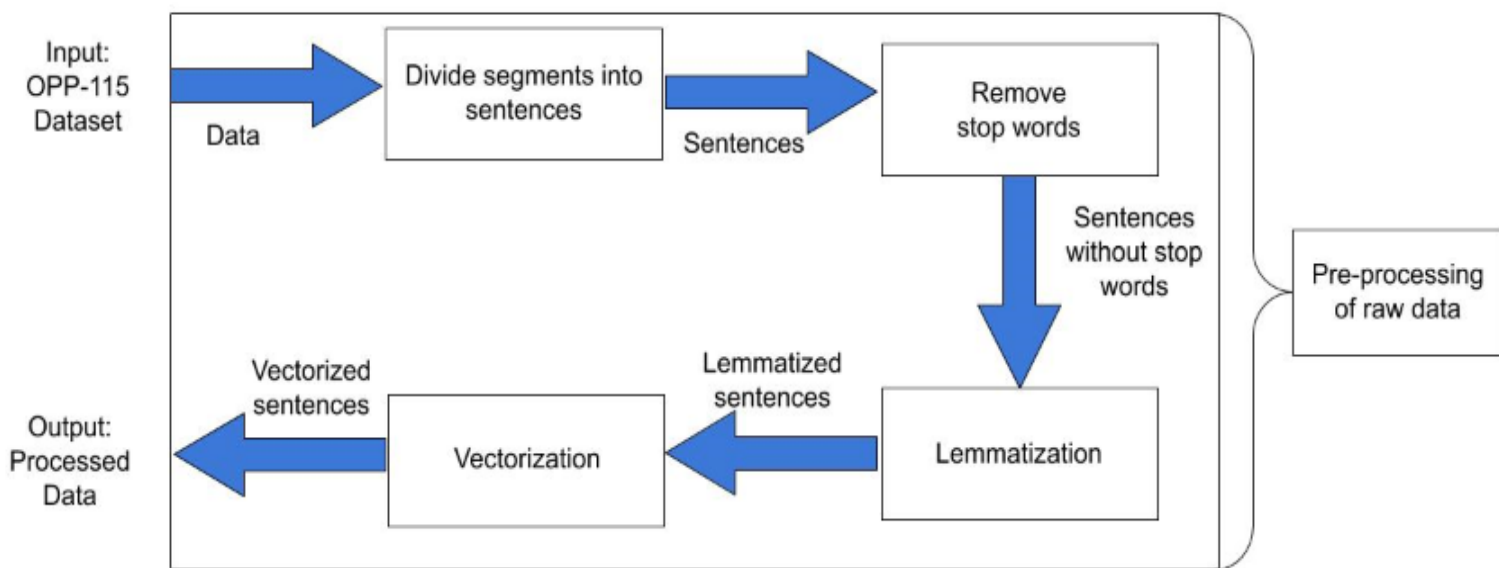


Figure 4.2.2.1.1 Pre-processing of raw data

as stop words: “a”, “an”, “the”, “are”, “is”, “what”, etc. Because these words give no information that can be used to classify the text, they are removed before any other action is taken.

The remaining text is then processed further using other widely known techniques like stemming and lemmatization. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. For example the word “studies” becomes “studi” or the word “studying” becomes “study”. On the other hand, lemmatization is usually referred to doing things properly with the use of a vocabulary and morphological analysis of words, this results in acquiring the base or dictionary form of a word. For example the words “studies” and “studying” become “study”. These techniques can be used individually or they can be combined. Finally, features are extracted from the segments. These features can be anything from individual words, to the number of words in a sentence or the number of characters in a segment. Anything that can help the classifier make a correct decision can be used as a feature. To determine what is best for our model we must first test it and see how it performs in each case. As seen above the technique used in this case was lemmatization (Figure 4.2.2.1.1). The reason for using lemmatization instead of stemming was the higher accuracy of the system, 74% and 70% respectively.

The next step is to convert the text into vectors that can be read and understood by our model. This too can be done in several different ways including term frequency-inverse document frequency (tf-idf), bag-of-words, etc. The best vectorization method for a given problem will also be determined by testing and comparing results.

The processed data is then divided into two parts, the training data and the testing data. The training data is given to our model to learn how to identify the class of each segment provided. The model learns by making a guess on what it thinks the correct answer is. It then makes changes on itself according to whether the prediction it made was correct or not. This process can be repeated multiple times

in a number of ways to find the optimal training requirements, but careful attention must be given not to overfit the model as this may result in problems later on.

After the model has been trained, the testing data is fed to it. Here, the model is faced with data it has not seen before and it is tested to see how it performs. If the model scores low it means that our training was not successful, If the model scores high in the training phase but scores low in the testing phase it might mean that our model has been overfitted and is only capable of recognizing data it has already seen. Depending on the testing and training results, the model's parameters need to be adjusted, as well as the pre-processing of the raw data. This is done in order to improve upon the model's performance as much as possible.

The process of adjusting may include things like: changing the size of the testing and training data, the number of iterations of the testing phase, the parameter of the model, etc. The pre-processing may change by adding more words in the list of stop words to be removed, as they have shown to provide no useful information. Additionally, trying different combinations of stemming and lemmatization or different vectorization techniques can result in vastly different results. Furthermore, a confusion matrix (Figure 4.2.2.2) can be used to find out what classes the classifier mostly guesses wrong and which classes confuse it the most. The x-axis shows the predicted class given by the classifier and the y-axis shows the correct class. The darker the colour, the higher the number inside the box. This information can help figure out why certain categories are being confused for others, which can help in improving the pre-processing further. Other data that can help with the differentiation of classes are the average length of each segment, the average word length, etc. For example Figure 4.2.2.3 shows the number of segments in each class, this information can be extremely useful because it may suggest that during the training of the classifier more positive reinforcement should be given for correctly guessing more rare classes as it won't have the chance to see them as often as others. Data like this, along with information about the classifiers performance as seen in Figure 4.2.2.4 which

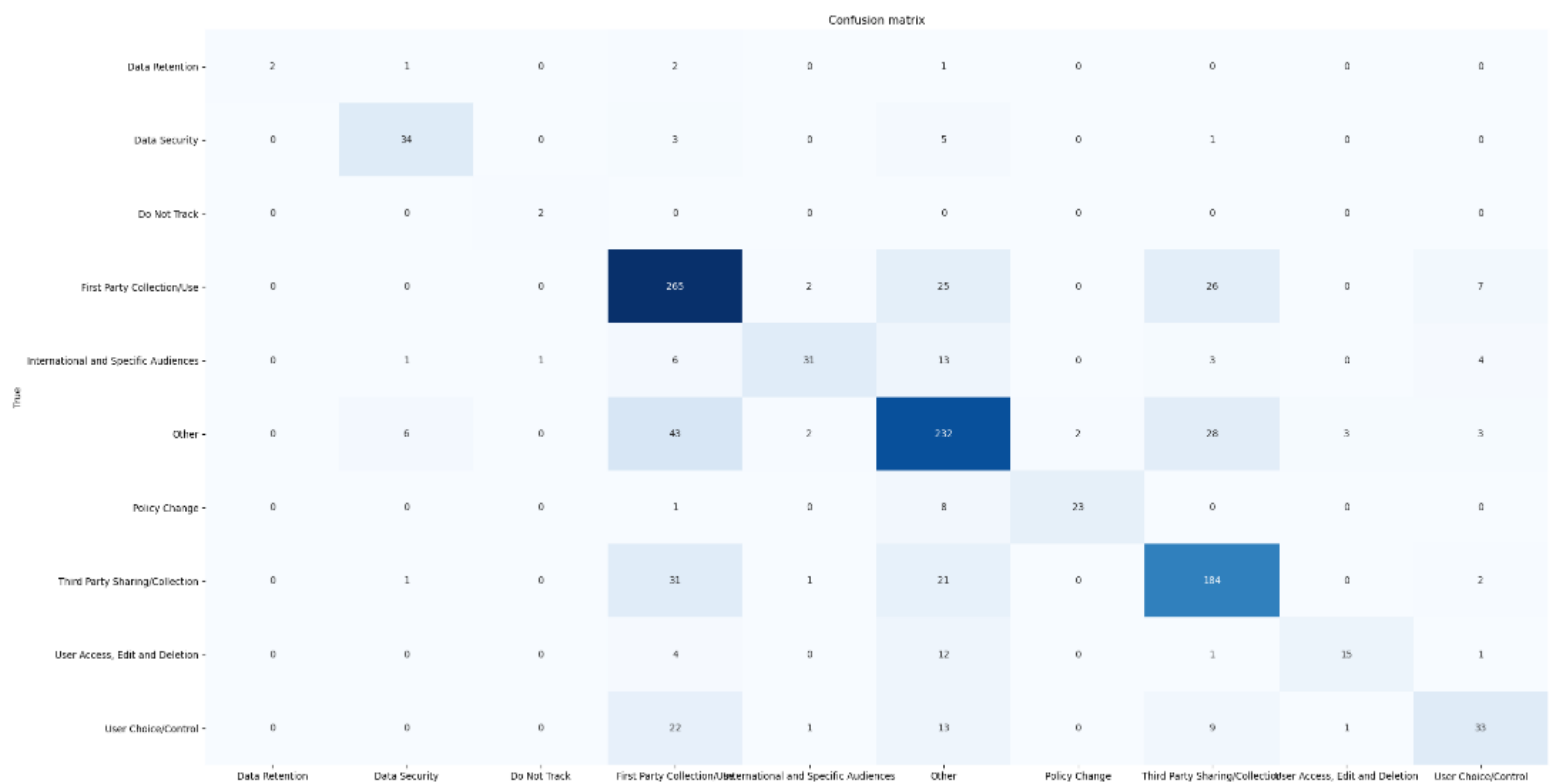


Figure 4.2.2.2 Confusion matrix for predictions (horizontal) and true values (vertical)

shows true positive rate - false positive rate (left) and precision - recall (right), can help determine what features should be used and what might need to be done to raise the classifiers performance. Graphs like the precision-recall curve in Figure 4.2.2.4, show that as the classifier keeps guessing, the precision decreases due to an increase in misclassification but the recall accumulates as more segments are being placed in the right class, the aim is to minimize misclassifications to keep the precision high while continuing to increase the recall. Additionally, a plethora of algorithms exist for creating the classifier itself, such as naive bayes, random forest, multilayer perceptron (MLP), support vector machine (SVM) and much more. These algorithms have their pros and cons and will perform differently depending on the problem they are tasked to solve. To find the optimal algorithm, testing is required to make an accurate comparison and find the right choice for the specific problem.

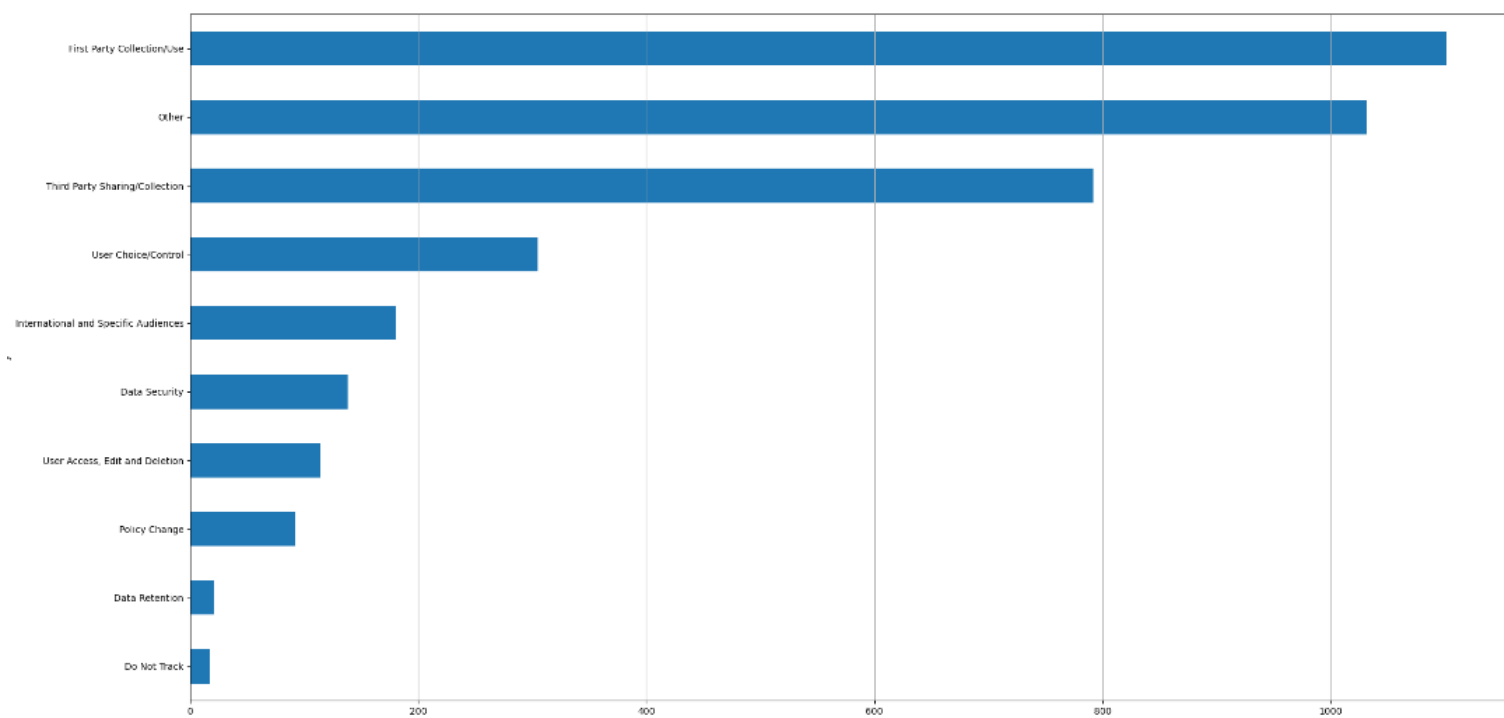


Figure 4.2.2.3 Class frequency in dataset

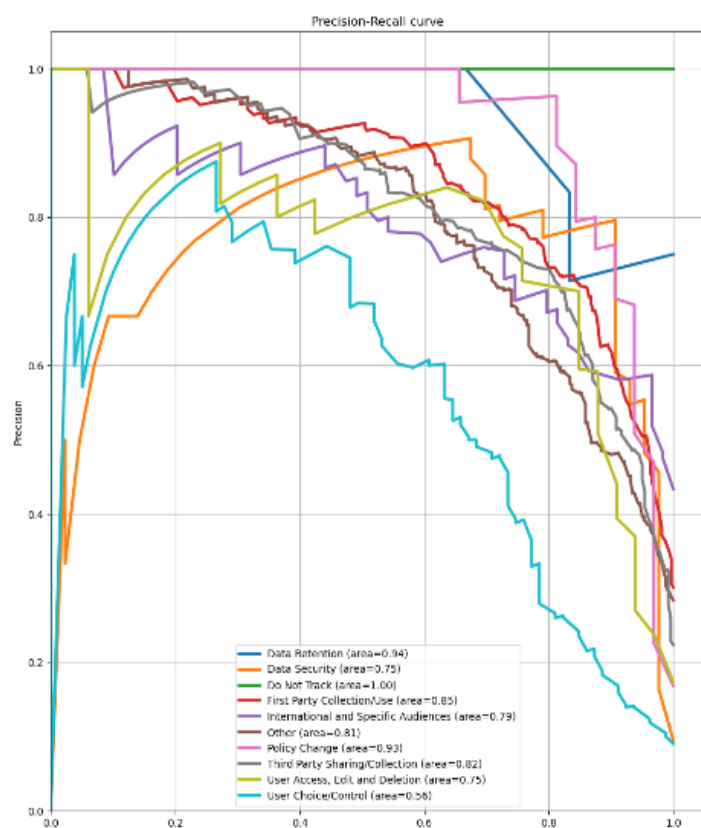
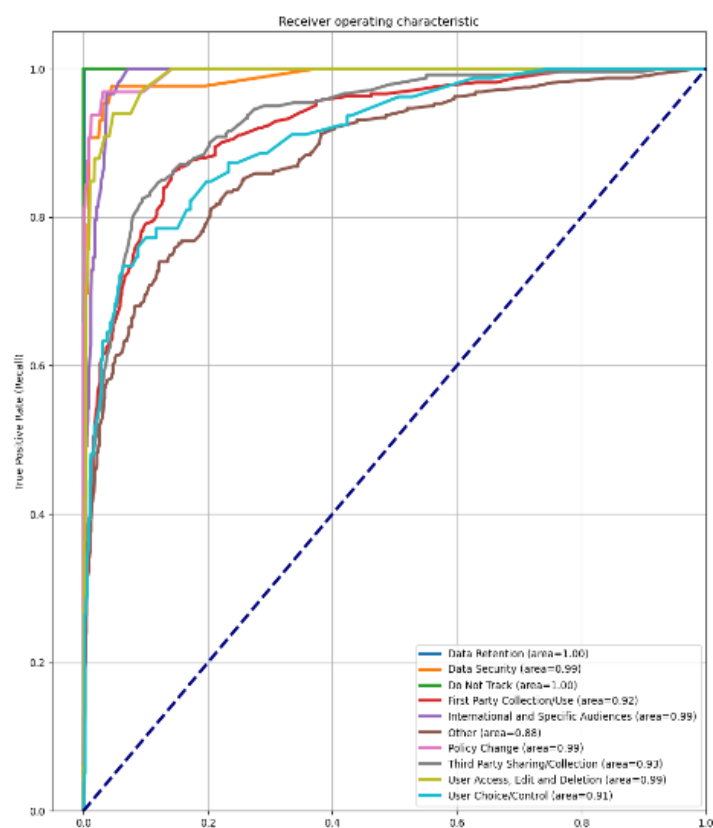


Figure 4.2.2.4 Receiver operating characteristic (left) – Precision-Recall curve (right)

4.2.3 Previous attempts

Before concluding on the use of a random forest classifier several other attempts were made to try and classify the text segments. The first attempt was made using a multilayer perceptron (MLP)([sklearn MLP classifier](#)). An MLP is a feedforward neural network which is composed of multiple layers of perceptrons as the name suggests. These perceptrons have a threshold activation and are interconnected transferring information to each other in order to produce a result. The reason an MLP was the first choice was because an MLP is guaranteed to converge on a solution even if that solution is not optimal, as long as the problem is linearly separable. After some testing it was clear that the results were not promising. The parameters that were changed during the training process were the number of hidden layers included in the MLP, hidden layers are the layers of perceptrons between the input and output layer, as well as the number of perceptrons contained in each layer. Additionally, different activation functions and solvers were tested with varying batch sizes and learning rates. After many attempts and various modifications/combinations to the parameters of the classifier, the accuracy did not exceed 50%. This is likely due to the fact that the amount of training data is not large enough and it does not have a wide range of diversity, as can be seen in Figure 4.2.2.3.

The second attempt was made using a [Naive Bayes classifier](#) algorithm, which is based on the Bayes Theorem with an assumption of independence among predictors. This means that the classifier assumes that the existence of a particular feature in a class has no relation to the existence of any other feature. The Naive Bayes classifier was chosen because it doesn't require as much training data as other classifiers, which is the main problem present in this situation. Additionally, it is highly scalable with the number of predictors and data points. It is fast and can make real time predictions. Unfortunately, despite the improvement in accuracy compared to the MLP, it was still not sufficient for this task. The reasons that may have caused the low accuracy, a maximum of 60%, may have been the zero frequency problem, where the algorithm assigns a zero

probability to a categorical variable, whose category in the test data set wasn't available in the training dataset. Furthermore, the basic idea of Naive Bayes which assumes that all features are independent, is rarely true in the real world, so that may have caused performance issues as well.

During these testing attempts, combinations of various other techniques were also used in order to see if there was a way to further improve performance. In these attempts different vectorization techniques were used as well as feature extraction methods. But despite the effort the improvements were not enough to make the classifiers as accurate as desired. Unfortunately, not all possibilities were tested due to the limited time available. Nevertheless, there may be possible combinations or other techniques that may have produced better results.

4.2.4 Final Classifier

Privacy Policy Beautifier uses a random forest classifier to separate the segments into the 10 classes mentioned before. Random forest or random decision forest is a supervised learning algorithm that fits or trains a set of decision tree classifiers. These decision trees are created during the training phase and each one outputs a prediction. All of the predictions are then taken into account and the best solution is selected by a means of voting. To implement the random forest classifier the library sklearn was used ([sklearn random forest classifier](#)). The variables of the classifier have been changed in order to improve the performance for this specific task. For example, the number of trees in the forest, the maximum depth of the tree, the number of features to be taken into account when looking for the best split, etc. All these variables can seriously affect the classifier in both positive and negative ways, this includes precision, recall and speed. That is why the training procedure of testing, adjusting and repeating is necessary to get the best possible results. After the model has been adequately trained and is capable of producing results with an acceptable accuracy and speed, it can be utilized in combination with the front end to create the web application that the end users will see and interact with.

4.3 Privacy Policy Beautifier

The web application was named “Privacy Policy Beautifier” to indicate its main purpose and functionality, which is to make privacy policies easier to read and more user friendly, or in short, more beautiful. The main issues with privacy policies as they are now, are their length, the vague language they use, and how hard they are to understand. This makes them tiring for users to read as they do not wish to spend the necessary time (about 201 hours a year [21]) or effort to go through a huge wall of text to find a piece of information that they may be interested in that they may not even be able to understand due to the use of vague or confusing words and phrases [17, 19]. Some of these issues are addressed by Privacy Policy Beautifier to encourage users to spend some time to read parts of the policy that may interest or worry them.

Implementation details. The frontend of the Privacy Policy Beautifier was built using Hyper Text Markup Language (HTML), Cascading Style Sheet (CSS), Javascript and [Bootstrap](#). Moreover, libraries such as [Google Charts](#) and [WordCloud2](#) were used for the visual representation of information. The structure, colors and design of the web application were selected in a way that would make the user’s experience as easy and pleasant as possible. Additionally, the [Flask](#) web framework was used for the creation of the web application as well as the creation of the APIs used to help the web application communicate with the backend. The Flask framework is written in python, as is the backend of the web application, and it is considered a micro framework because it does not need any particular tools or libraries to function like you would see in other frameworks.

The user can use the web application by inserting the URL of an HTML page containing a privacy policy in the provided box as seen in Figure. 4.3.1, or by pasting the whole or part of the privacy policy in the “Policy Full Text” tab, as shown in Figure. 4.3.2. In addition, the user can view more information about the page, as well as instructions on how to use it, and contact info, in the “General Information” tab as seen in Figure 4.3.3.

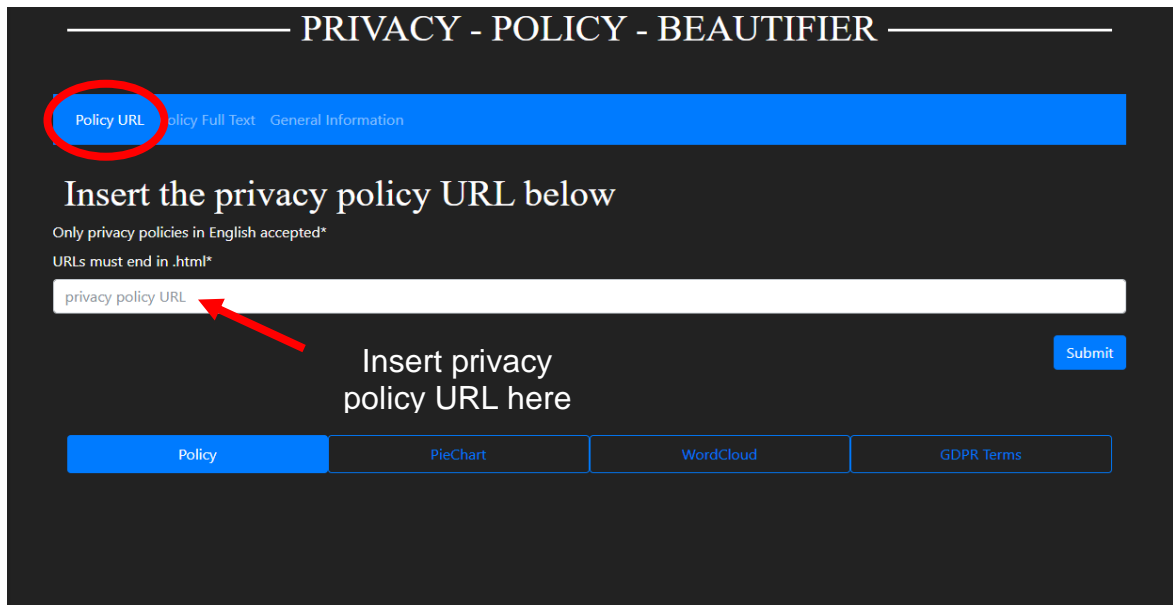


Fig. 4.3.1 Privacy Policy Beautifier “Policy URL” tab

Afterwards, the user can press the submit button or the “enter” key so that he/she may receive the beautified version of that information. When the user’s request has been submitted the policy URL or text is sent for processing. If the user has submitted a URL, an extra step has to be taken before any analysis can take place. The URL is used to load the HTML file, which is then given to a parser in order to remove any unwanted tags or information that are not needed. Next, the parser will output the text that needs to be analysed, so the same process is applied from here for both cases (URL or text submission). Firstly, all stop words

are removed from the text. The list of stop words to be removed contains the standard stop words from the nltk library, which contains 127 words. Moreover,

The screenshot shows the 'PRIVACY - POLICY - BEAUTIFIER' interface. At the top, there is a navigation bar with three tabs: 'Policy URL', 'Policy Full Text', and 'General Information'. The 'Policy Full Text' tab is selected and highlighted with a red circle. Below the navigation bar, the main heading is 'Insert the privacy policy text below', followed by a sub-note: 'Only privacy policies in English accepted*'. A large white text input area contains the placeholder text 'privacy policy'. A red arrow points from the text 'Paste text from a privacy policy here' to the input area. To the right of the input area is a blue 'Submit' button. At the bottom, there is a row of four buttons: 'Policy', 'PieChart', 'WordCloud', and 'GDPR Terms'.

Fig. 4.3.2 Privacy Policy Beautifier “Policy Full Text” tab

The screenshot shows the 'PRIVACY - POLICY - BEAUTIFIER' interface with the 'General Information' tab selected and highlighted with a red circle. The main heading is 'What the site is about:', followed by a paragraph: 'This page is designed to help make privacy policies a bit more friendly and easy to read.' Below this is another paragraph: 'Usually a privacy policy is a massive wall of text that the user can not be bothered to read as it contains a lot of information they don't care about.' This is followed by a third paragraph: 'This site will help you find and focus on what is more important to you as a user so you know what you are getting into every time you press that infamous "i have read and accept the terms and conditions" button.' The next heading is 'How to use:', followed by a list of instructions:

- Choose how to insert the privacy policy you want to beautify
 - **"Policy URL:"** Here you can place the url of the site where the privacy policy is located. Note: some websites prevent automatic inspections of their sites so this method may not always work. If you are presented with an error message please try using the Policy Full Text option.
 - **"Policy Full Text:"** Here you may paste the text of the policy you want to beautify. It may be a portion of the policy or the entire text.
- After inserting the privacy policy click the **"submit"** button
- Below the "submit" button you have options on how to see the beautification results.
 - **"Policy:"** presents the privacy policy divided and color coded into 10 different categories that can be seen on the side of the text. The categories can be clicked which will magnify all sections in that category and minimize all the rest to help you find what you are looking for.

Figure 4.3.3 Privacy Policy Beautifier “General Information” tab

words that were found to not contribute or provide any valuable information for the model are also added to this list. In this stage, the stemming and lemmatization processes are executed to further improve the performance of the classifier. Afterwards, the text is divided into segments, which are given to the model for classification. Each segment is evaluated and assigned a class by the trained classifier and the results are sent back to be handled and displayed. This process can be seen in Figure 4.3.4. At the same time the text is analysed in order to find references to the GDPR. The results of this analysis are also sent back to be displayed to the user under the “GDPR Terms” tab, as seen in Figure 4.3.5. This functionality was based on the results of Γιώργος Ζαμπά [1].

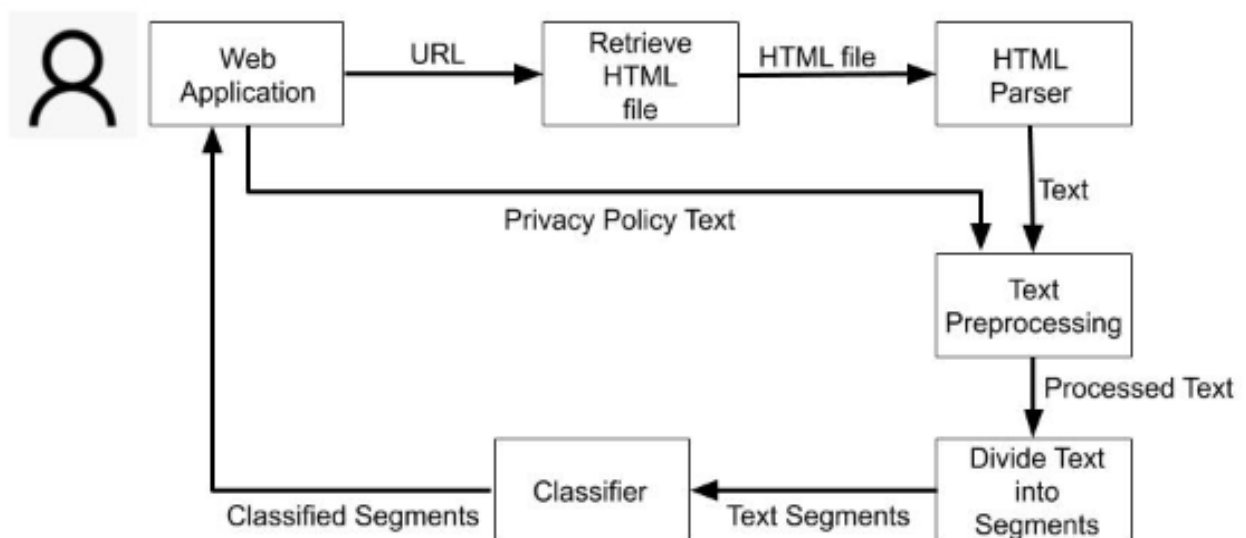


Figure 4.3.4 Process summary of the entire system

The main feature of the web application is located in the “Policy” tab. Here, 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, this is seen in Figure 4.3.6. The color of the segment represents the class it belongs to. The colour associated with each category can be seen on the left side of the screen, for example white represents the category “Other” and red the category “Data Security”. By pressing the button of a specific category, the text

that belongs to that category becomes bigger, the rest of the text becomes smaller, and the page scrolls to bring into view the first appearance of the selected category, as seen in Figure 4.3.7. This is done to help the user find and read the parts of the privacy policy that he/she is interested into. To revert this effect, the user can click on the selected category again or click the clear filter button below the categories, by doing this the text turns back to normal.

PRIVACY - POLICY - BEAUTIFIER

[Policy URL](#)
[Policy Full Text](#)
[General Information](#)

Insert the privacy policy URL below

Only privacy policies in English accepted*

URLs must end in .html*

privacy policy URL

Submit

Policy

PieChart

WordCloud

GDPR Terms

Lawfulness of Processing	
Lawfulness of Processing	NO
Consent	NO
Contract	YES
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
To Erase Your Data	NO

Figure 4.3.5 Privacy Policy Beautifier “GDPR Terms”

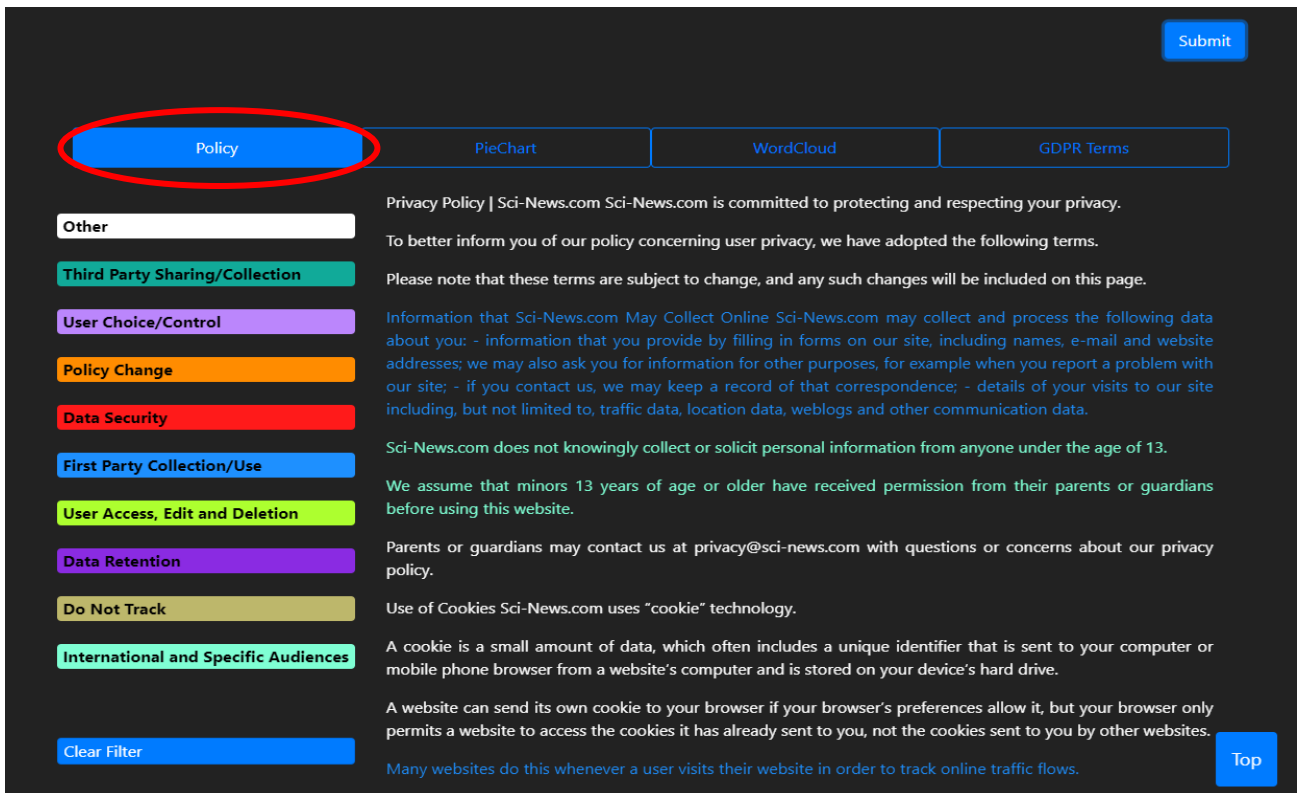


Figure 4.3.6 Privacy Policy Beautifier "Policy" Text display option

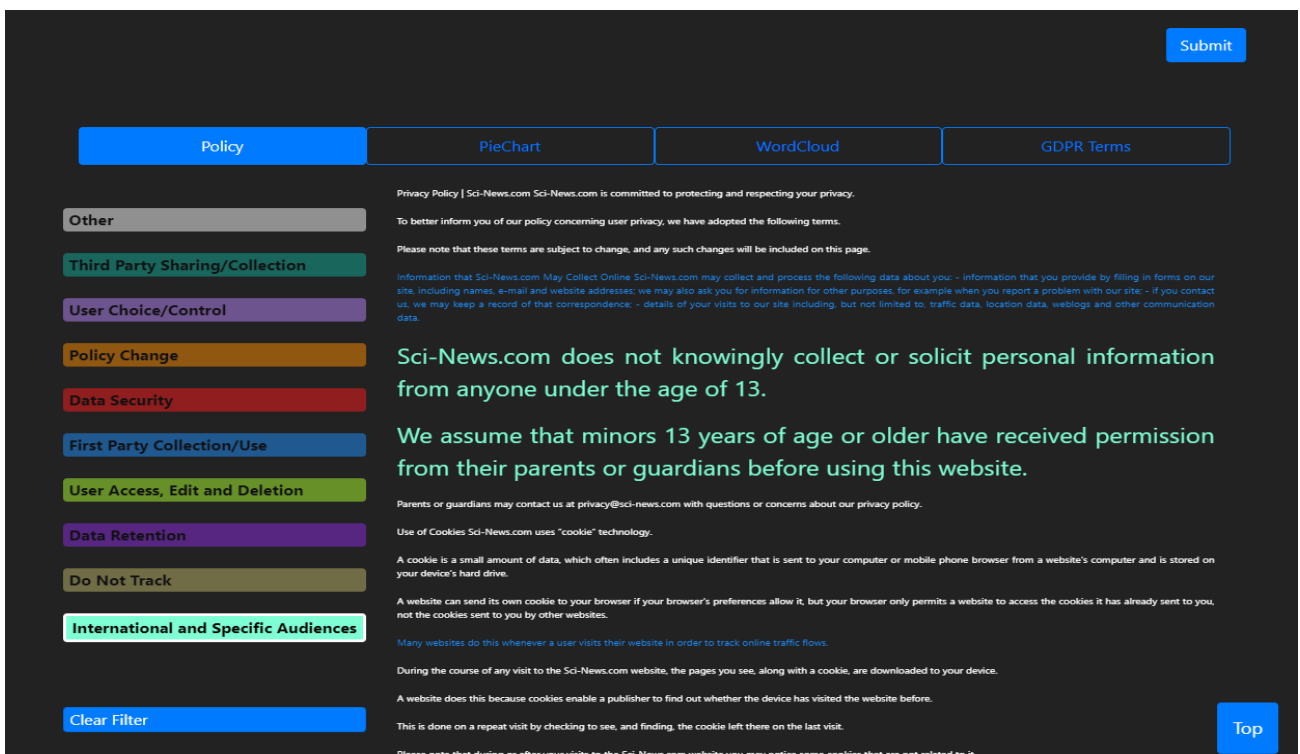


Figure 4.3.7 Privacy Policy Beautifier "Policy" Text display option with a filter selected

The classified segments given by the classifier are also used in the “PieChart” tab. Here a pie chart is created to show the percentage of each class that the privacy policy contains, as seen in Figure 4.3.8. This is calculated based on the number of segments each class has compared to the total number of segments in the privacy policy. The pie chart is to help the user see where more emphasis was given in the privacy policy and whether or not a certain category is present or not. It is a clear and easy way to summarize the contents of the privacy policy that requires very little time from the user to read and understand.

The final tab titled “WordCloud”, presents to the user a word cloud, as the title implies, with the most frequently appearing words in the privacy policy represented in a larger font and in the center, as seen in Figure 4.3.9. The less frequent a word is the smaller it appears in the word cloud. This representation was included to give the user an idea of what is being said in the privacy policy and how much they are expected to see it without having to go through the entire text.

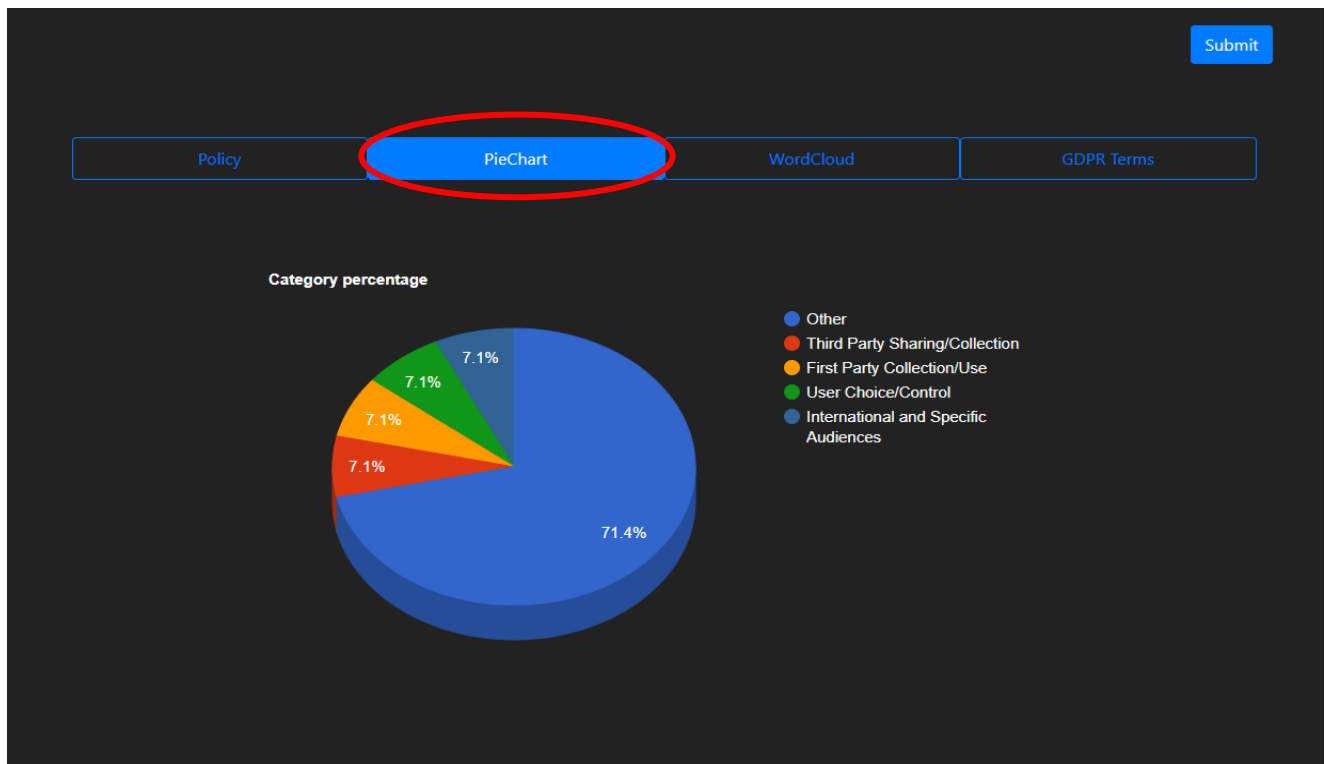
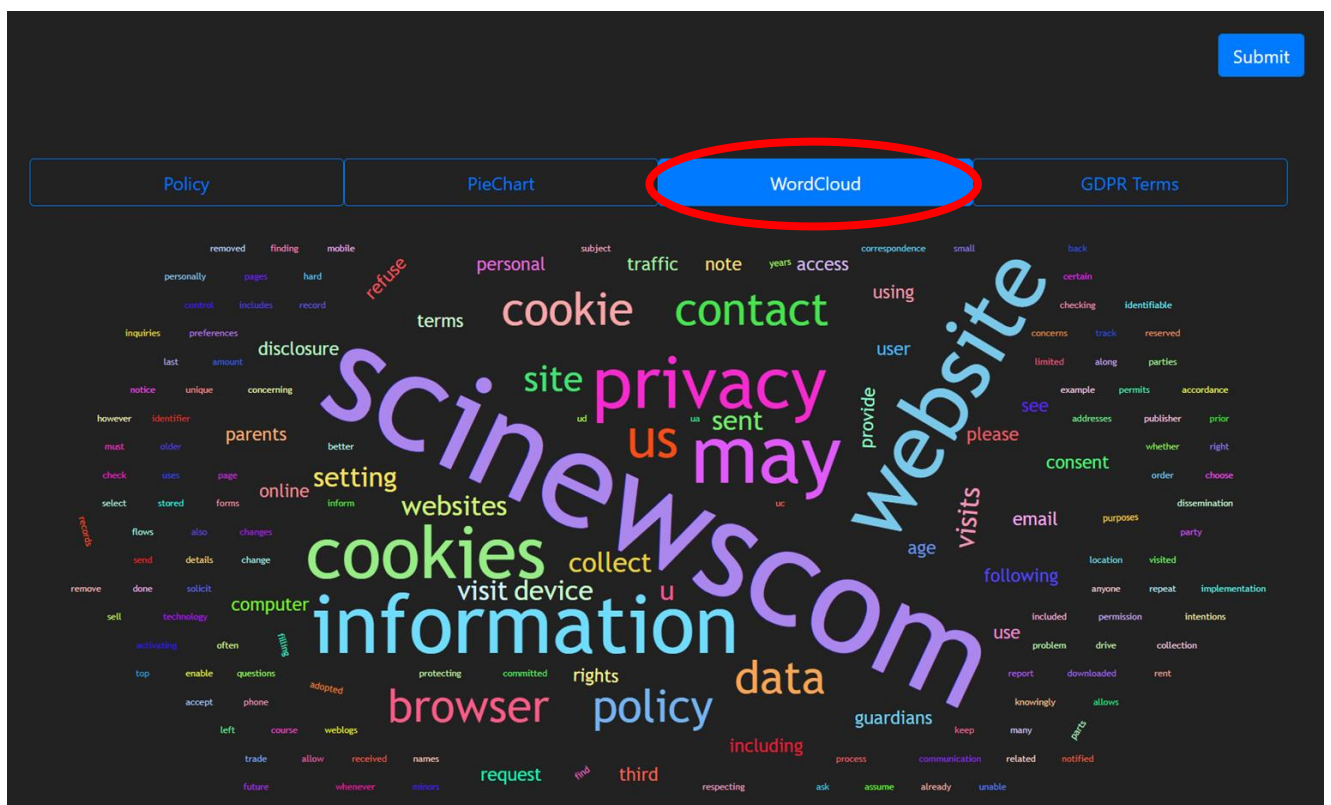


Figure 4.3.8 Privacy Policy Beautifier “PieChart” display option



Chapter 5

Platform Evaluation

5.1 Introduction	53
5.2 Classifier Evaluation	54
5.3 User Evaluation	56
5.3.1 The Questionnaire	56
5.3.2 Evaluation results	58

5.1 Introduction

This chapter focuses on the evaluation of (i) the classifier used to classify the different segments of the privacy policy into categories, and of (ii) the web application itself, which is what the end users will see and interact with.

Evaluating the classifier is important because most of what the Privacy Policy Beautifier does relies on its output. This means that if the results given by the classifier are wrong, then the users will be provided with false or misleading information. Such scenarios are best to be avoided or at least reduced as much as possible. Additionally, it was made to help users raise their privacy awareness, which can't be achieved by feeding them false information.

Furthermore, evaluating the web application itself is even more important, since it was designed to encourage users to read fully or partially the privacy policies they encounter. If the web application is not user friendly or not useful to the

users, then they will not use it. If no one uses the web application then it makes no difference how accurate the classifier is. To make sure the web application is user friendly a questionnaire was created and given to users. This questionnaire was used to extract general information about the experience users have had with privacy policies in the past, and to see how they felt when using the Privacy Policy Beautifier. If the users had a more pleasant experience with the Privacy Policy Beautifier rather than with the original privacy policy, then the web application has achieved its goal. Additionally, having the users feedback can help improve the web application by adding, removing or changing certain features.

5.2 Classifier Evaluation

Evaluating the classifier is very important in creating a usable and reliable tool to help users raise their privacy awareness as mentioned above. This is why the classifier underwent rigorous testing during its creation and every step of the way. The testing phase always followed the training phase to see whether the classifier was able to learn and be able to correctly make predictions concerning the categorisation of segments from the text provided. Testing the classifier after training helps figure out if changes need to be made to the training phase, or if changes that have been made to the training phase were beneficial to the classifier or not. This process of training, testing, adjusting and repeating is aimed at increasing the classifier's accuracy. The training and testing of the classifier were both done using the OPP-115 dataset [31].

There are several ways to determine the accuracy of a classifier, these include recall, precision and f1 score. Recall is the fraction of relevant documents or items that are successfully retrieved. For example, if the classifier classified 80 items as "Data Security" and the items labeled "Data Security" in the testing data was 100, then the classifiers recall was $80/100 = 0.8 = 80\%$. Precision is the number of correctly classified items given by the classifier. For example, if the classifier's output indicates that 90 items belong to the class "Policy Change", but out of those items only 65 were actually of that class, then the classifier's precision is

$65/90 = 0.72 = 72\%$. The f1-score is used to combine these two metrics into one, and is needed as it is easy to achieve a high recall or precision and claim an accurate system when in reality it may not be. The mathematical equation used for the f1-score is $F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$. This formula can be altered to favor either the precision or recall of a system. In this case the F1 formula was used in its original form.

While evaluating the classifier, a recall, precision and f1-score was calculated for each class, and the average of the f1-scores were used as the accuracy score for the classifier. The highest accuracy achieved by the classifier was 74% and that is the model used by the Privacy Policy Beautifier. Even though the recall and precision scores were not directly used to determine the classifiers accuracy, they were used to pinpoint potential issues in the training phase that were then addressed and corrected in the adjusting phase. The score of this classifier may not be as high as other like the one found in Polisis [10], which has an average score of 88.4% but it is on par with others like the one in PrivacyCheck [32] with a score of 40%-73%.

Detail:					
	precision	recall	f1-score	support	
Data Retention	0.67	0.33	0.44	6	
Data Security	0.77	0.80	0.79	41	
Do Not Track	1.00	0.83	0.91	6	
First Party Collection/Use	0.73	0.81	0.77	337	
International and Specific Audiences	0.60	0.45	0.51	47	
Other	0.70	0.80	0.75	319	
Policy Change	0.88	0.78	0.82	27	
Third Party Sharing/Collection	0.79	0.69	0.74	251	
User Access, Edit and Deletion	0.76	0.53	0.63	30	
User Choice/Control	0.83	0.58	0.68	74	
accuracy			0.74	1138	
macro avg	0.77	0.66	0.70	1138	
weighted avg	0.74	0.74	0.74	1138	

Figure 5.2.1 Classifier scores separated by category including an overall accuracy score

5.3 User Evaluation

The most important part of designing the Privacy Policy Beautifier was to make it as user friendly as possible. This means it had to be easy to understand, easy to use and to provide a pleasant user experience. Additionally, it had to be capable of providing users with correct and reliable information in a way that would help them increase their privacy awareness and encourage them to spend some time reading privacy policies they encounter. The accuracy and validity of the information given by the Privacy Policy Beautifier was analysed in the previous section (5.2 Classifier Evaluation). In this section, an evaluation of the user experience will be analysed. More accurately, a more in depth explanation will be given on how users were able to evaluate the Privacy Policy Beautifier, as well as their experience with other privacy policies in the past.

The feedback and information provided by the end users is very valuable as it helps make changes and improvements to the design and implementation of the Privacy Policy Beautifier. This is done by seeing whether users liked or disliked a particular section or feature or the web application. If users are more drawn to a particular feature then it might be beneficial to emphasize and improve it even more. On the other hand, if users seem to dislike or not understand a certain feature then it can be either improved, changed or even removed entirely to make sure users don't waste time using it. Finally, other changes or new features may be recommended by the users that were not originally thought of.

5.3.1 The questionnaire

To gather as much feedback from users, a questionnaire was created and given to users. This questionnaire was made as short as possible to not bore or discourage users from answering it. It is made up of 9 sections, each dedicated to a specific topic or aspect of interest. The questionnaire was answered by 89 different individuals, with a variety of ages and educational backgrounds. It was important to include as much variety as possible due to the fact that all people encounter privacy policies, despite their gender, age, or level of education. In

addition, it gives a more complete overall idea of how people feel or perceive privacy policies.

The first section of the questionnaire informs the participant of the questionnaire's purpose and asks whether or not they consent to be a part of this survey. If the participant does not consent, then they will not move on to the following questions.

The second section focuses on gathering basic information from the participant. This information includes their age, their gender if they wish to disclose it, whether or not they consider themselves to be a technical expert and the level of education they have completed or are currently attending. Next, if the participant has completed or is attending a bachelor degree or higher they are taken to section 3 where they are asked to disclose what their educational background is, if they wish to do so, otherwise they are taken directly to section 4. This general user information will be useful in reaching conclusions later on.

In section 4, the participant is asked whether or not they have read a privacy policy before from any website or application. This is important to know, as many users tend to completely avoid reading privacy policies, which is worrying in our day and age. If the participant answers “No” they are taken to section 7, otherwise if they answer “Yes” or “Partially” they move on to section 5.

In section 5, the participant is asked 3 questions regarding their experience while reading privacy policies in the past, as well as if they would ever consider reading one again in the future. If the participant says they would not read another privacy policy in the future, they are taken to section 6 where they are asked to elaborate on why they would not read another privacy policy again.

Next, in section 7 the participant is asked what would make them read a privacy policy and what would deter them from doing so. These questions were made in

order to let the participant express themselves freely, and to see their concerns when it comes to privacy policies or their biggest complaints about them.

In section 8, the participant is asked to select or write the reasons that made them not wish to read a privacy policy and they are provided with a selection of options as well as the option to add extra reasons that may not have been included. This section is very important as it provides even more insight on the perception of privacy policies to the users.

The final section of the questionnaire, section 9, asks the participant to visit the Privacy Policy Beautifier and use it before continuing on. This section's sole purpose is to evaluate the experience the participant had while using the Privacy Policy Beautifier. The answers gathered here are the key to determine whether the Privacy Policy Beautifier is doing the job it was created to do, and how it can be improved in future updates.

The questionnaire can be found in Appendix A and in the following link: <https://forms.gle/MBCyuRPptHeqbVjh8>

5.3.2 Evaluation Results

The questionnaire was answered by a total of 89 people of an age range from 18 to above 60. The majority of participants were between the ages of 18 and 24 (41.6%), as seen in Figure 5.3.2.1, and more than half are male (56.2%), as seen in Figure 5.3.2.2. Despite the fact that the majority of responses came from younger ages, there was a wide variety of age groups, which can provide an overview that is closer to reality. Most participants claim to have some technical knowledge (40.4%), but only a minority claim to be a technical expert (25.8%) (Figure 5.3.2.3). Additionally, almost half of the participants have or are currently attending their bachelor degree, as seen in Figure 5.3.2.4. When it comes to the participants' background education, there is a very wide spectrum ranging from Computer Science, to Agriculture, to Law. Surprisingly, from all 89 responses,

What is your age?

89 responses

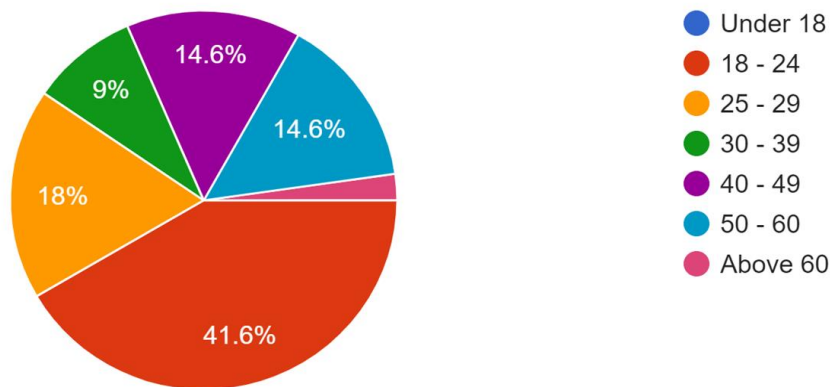


Figure 5.3.2.1 Age groups of participants

Gender

89 responses

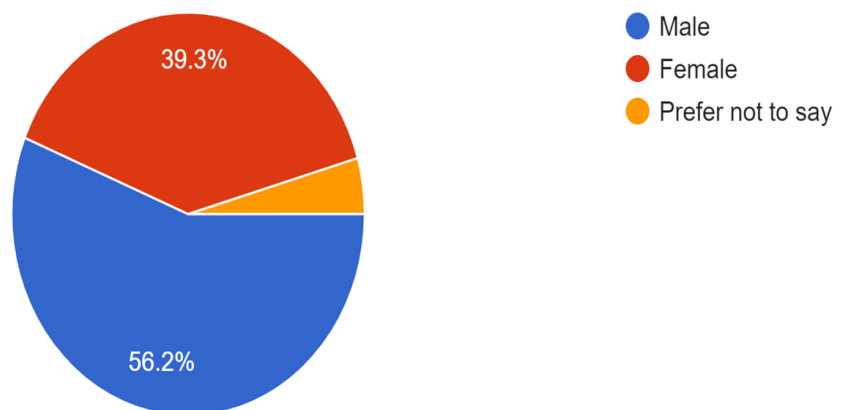


Figure 5.3.2.2 Gender groups of participants

only 15 people have never read a privacy policy before. The rest have either read one completely (29.2%) or at least partially (53.9%) (Figure 5.3.2.5). This indicates that people are spending time and effort to go through privacy policies, which is a good sign that they are interested in finding information concerning their privacy. Unfortunately, from the 74 people that have read privacy policies before more than half (45 responses) did not have a pleasant time doing so and

Would you say you are a technical expert?

89 responses

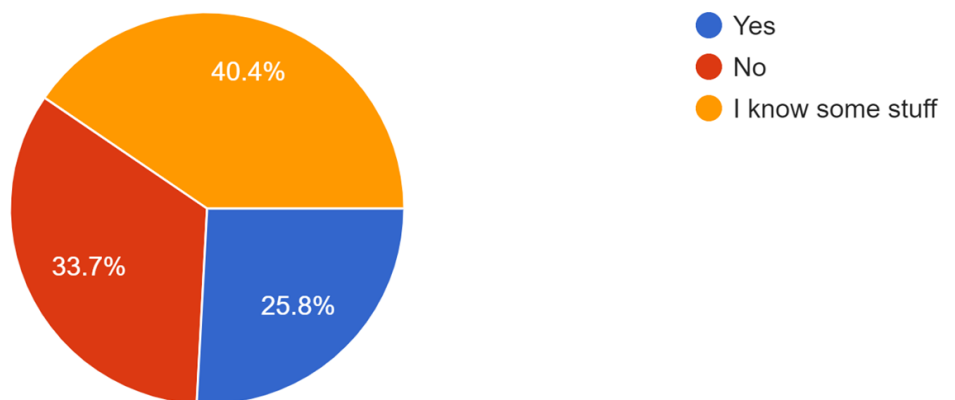


Figure 5.3.2.3 Technical experts - people with some technical knowledge - people with no technical knowledge

What level of education have you completed or are you currently attending?

89 responses

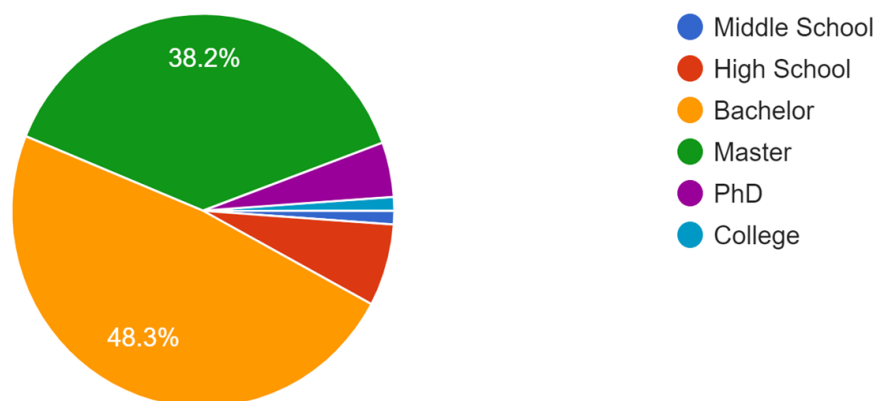


Figure 5.3.2.4 Levels of education

had a pretty hard time finding what they were looking for (40 responses) (Figure 5.3.2.6). These results indicate that privacy policies remain somewhat unpleasant to users and are still hard to navigate for most. But despite that, it is not a dire situation, as 23 responses say they were neither pleased or displeased with their experience, and 16 of responses were able to find what they were looking for easily (Figure 5.3.2.6).

Have you ever read a privacy policy from any website or application?
89 responses

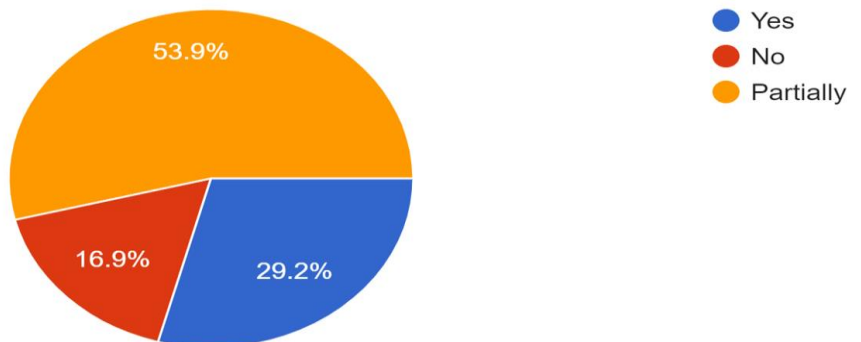
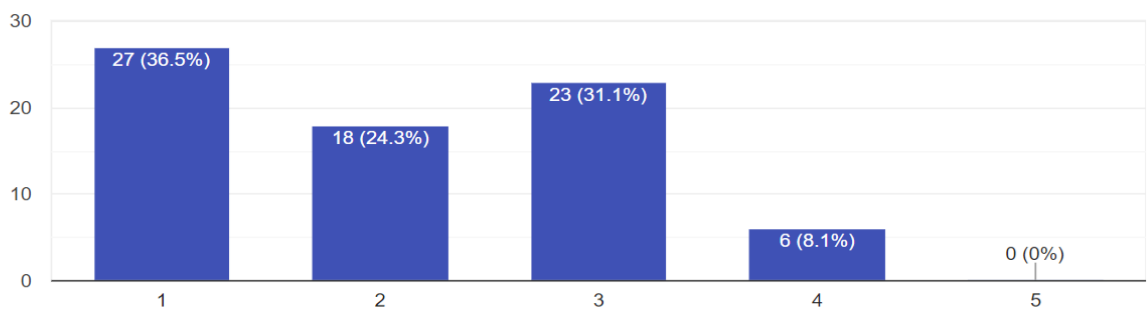


Figure 5.3.2.5 Users that have read privacy policies

Read a privacy policy

Was your reading experience enjoyable?

74 responses



Did you find the information you were looking for easily?

74 responses

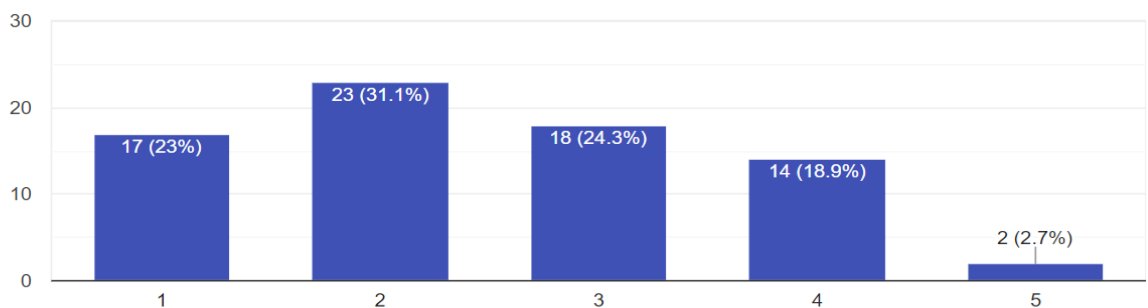


Figure 5.3.2.6 Previous experience with privacy policies

Fortunately, only a minority of people would not consider reading another privacy policy (16.2%), whereas most of them would either consider reading (60.8%) or would definitely read another one (23%) (Figure 5.3.2.7). The reasons people gave as an answer for not wanting to read a privacy policy were more or less expected, with responses saying that privacy policies are too long, too complicated, or contain too much information and are a waste of time.

Would you ever read another privacy policy?

74 responses

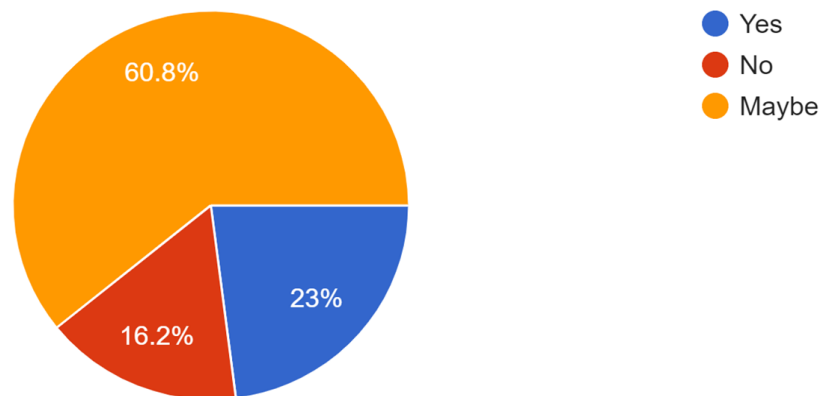


Figure 5.3.2.7 Users that would read privacy policies

Similarly, the people that would read another privacy policy in the future said that the things that would discourage them from doing so are the massive length of text, if it was too difficult to understand (using complex vocabulary or terminology) or if it takes too much time. This result can be seen again where the top 3 reasons people said made them not want to read a privacy policy in the past were (in decreasing order): too long, time consuming or too hard to understand/confusing (Figure 5.3.2.8). What seems to be worrying is the large number of people who answered “I didn’t care” which indicates a lack of interest in increasing one’s privacy awareness.

What reasons made you not want to read a privacy policy in the past?

89 responses

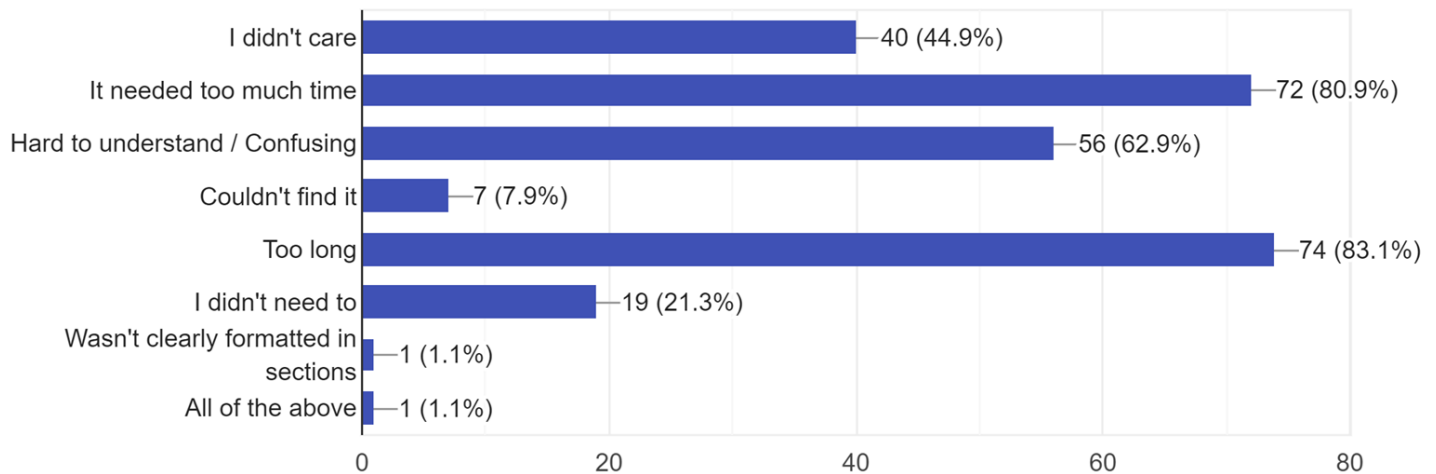


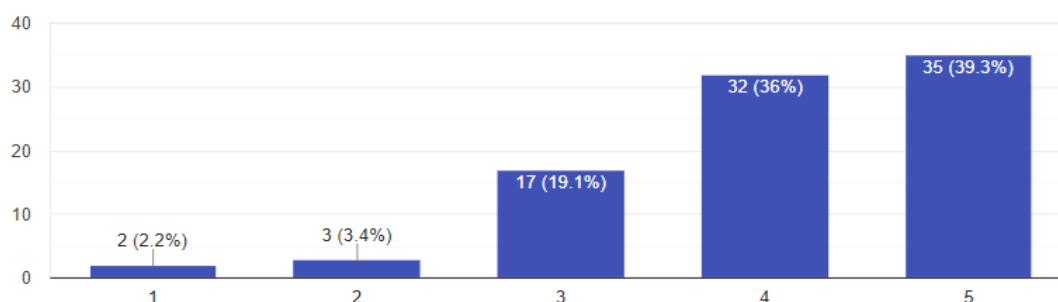
Figure 5.3.2.8 Reasons for not reading privacy policies in the past

The most important part of the questionnaire was section 9, which was focused on the experience users had while using the Privacy Policy Beautifier. The users were asked whether they found the web application easy to use and whether or not it made the privacy policy easier to read. These two points were the main goal since the inception of this project. From the responses it is clear to see that the majority of users found the Privacy Policy Beautifier relatively easy to use and it made privacy policies at least a bit easier to read than their original form (Figure 5.3.2.9), this metric can also be affected by the privacy policy given as some may be more user friendly than others.

As seen in Table 5.3.2.1 users that claim to be technical experts found the Privacy Policy Beautifier easier to use than users that claim to not be technical experts or that have only some technical knowledge. Furthermore, users that are technical experts are more likely to use Privacy Policy Beautifier in the future as seen in

Did you find it easy to use the privacy policy beautifier website?

89 responses



Was the privacy policy easier to read?

89 responses

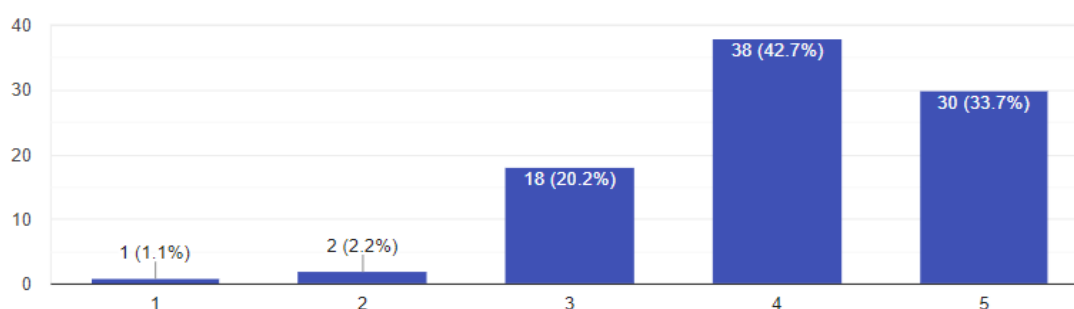


Figure 5.3.2.9 Experience while using the Privacy Policy Beautifier

Table 5.3.2.2. Moreover, it appears that users that are technical experts prefer the textual and GDPR representations, 18 out of 23 and 13 out of 23 respectively but they don't seem to like representations like the pie chart (6 out of 23). In contrast, non-technical users seem to prefer the pie chart representation as more than half chose it (16 out of 30) but not a smaller percentage of them chose the textual or GDPR representation, 13 out of 30 and 9 out of 30 respectively.

		Did_you_find_it_easy_to_use_the_PrivacyPolicyBeautifier					Total
		1	2	3	4	5	
Would_you_say_you_are_a_technical_expert	I know some stuff	2	1	8	12	13	36
	No	0	2	7	10	11	30
	Yes	0	0	2	10	11	23
Total		2	3	17	32	35	89

Table 5.3.2.1 How easy was it to use the Privacy Policy Beautifier for technical and non-technical users

		Would_you_consider_using_privacy_policy_beautifier_again			Total
		Maybe	No	Yes	
Would_you_say_you_are_a_technical_expert	I know some stuff	11	4	21	36
	No	9	2	19	30
	Yes	5	1	17	23
Total		25	7	57	89

Table 5.3.2.2 Would users consider using Privacy Policy Beautifier again. A comparison between users that are technical experts and those who are not

Also, it was very important to know whether users would consider using the Privacy Policy Beautifier again to find out information they wanted from privacy policies. If the web application was appealing enough for users to revisit regularly every time they encounter a privacy policy they want to read, then it would mean the project was a success, and people would spend time and effort in raising their privacy awareness. Promisingly enough, more than half of users said they would consider using Privacy Policy Beautifier in the future (Figure 5.3.2.10).

Furthermore, it was very important to know what features users liked the most, in order to figure out what is more appealing to the general public and focus more on improving them. As seen in Figure 5.3.2.11, most users preferred the textual representation of the privacy policy, but also liked the pie chart and the GDPR representation respectively. In addition, users seemed to like the word cloud representation the least by quite a margin. This means that it might need significant improvements or it might be better to remove it so as to not waste the users' time or even confuse them.

Finally, the users were asked if they had any comments or suggestions for future versions of the Privacy Policy Beautifier. Some suggestions have already been added, like being able to disable a filter by clicking it again instead of having to press the "clear filter" button, to scroll till the first instance of the category the user clicked on is visible, or making the selected filter more clear and distinct.

Many of the users' comments were positive or neutral as seen in Figure 5.3.2.12, which is very encouraging for the web applications development. Additionally, users had some suggestions about new features that they would like to see like the summarization of the text or topics, the introduction of a word search or the ability to apply multiple filters at once (Figure 5.3.2.13). Moreover, users also suggested small changes or adjustments that they believe would make their experience more enjoyable or the web application more user friendly. These suggestions can be seen in Figure 5.3.2.13 and some of them include the use of different colours that are more distinct from each other, the use of tags for further explanation of the categories and much more. All these comments and suggestions are encouraging as it shows that there is room for improvement as well as that users are paying attention and are involved to help a tool that can one day be part of their daily life improve.

Would you consider using privacy policy beautifier again?

89 responses

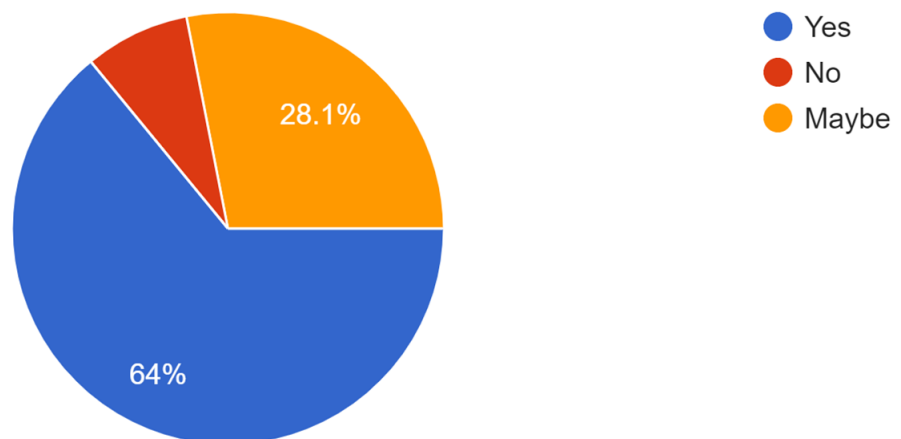


Figure 5.3.2.10 Would users use Privacy Policy Beautifier again

Which presentation/s of the content of the privacy policy did you prefer?

89 responses

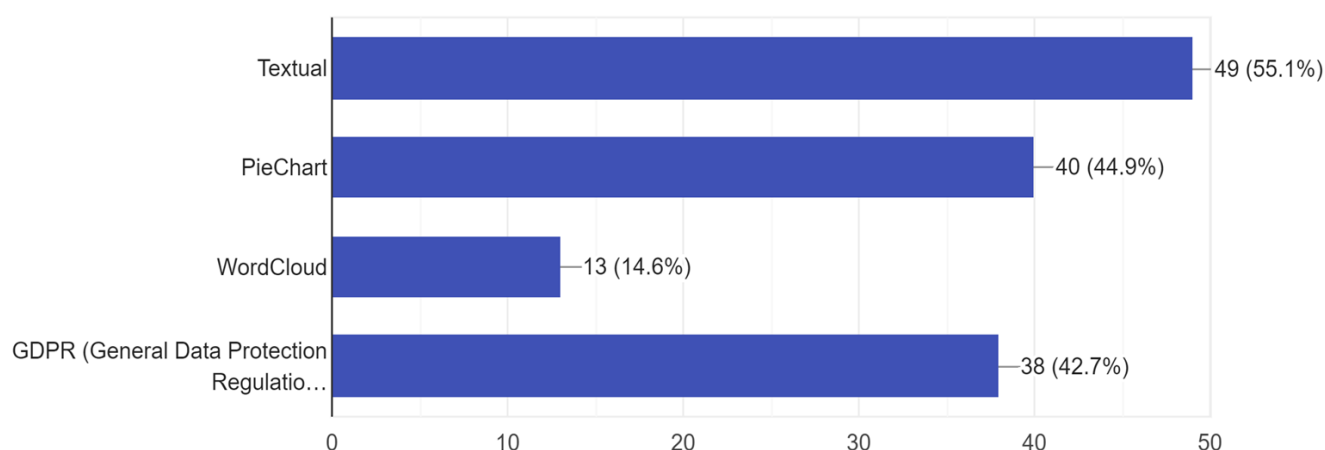


Figure 5.3.2.11 Preferred method of displaying information

Satisfied	Comments
No, everything was perfect. :)	Privacy policies will always be boring to read
I believe it is a great tool that can help users finding the information they need	Explain to people that even if they copy an address that does not end with html, all they need to do for the beautifier to work is to type an html at the end.
overall a nice site, never thought of the alternative ways to present privacy policy	In the wordcloud i think different size and colours of the words-appeared may confuse the reader and make difficult to catch the point
Is it perfect. Any change.	No
Its perfect	no
Easy to answer and to understand	No
It made understanding policies much simpler.	no
Very nice presentation and fast response	
No I think it was pretty straight foreword	
Was pretty good	
Very nice	
No it is very good !	

Figure 5.3.2.12 Positive or Neutral or Suggestive comments from users

Complaints	Small changes / Adjustments	WordCloud	New Features	Further Explanation
The output of the text in the Policy tab is inconsistent. Some text is too big, others are too small.	I would prefer colors that aren't similar with each other for each category	The word cloud is not useful	If it could give me a summary of the section of interest	Small explanations about the different things in pie chart and textual window for the people that have no clue about the topic
Difficult to find privacy policy links strictly ending in .html.	after pressing button it would be nice to be transferred automatically to that highlighted section. it would be nice if the categories that are not covered in the privacy policy were put in a separate section so that the user does not click on button that do nothing OR if the button does nothing display a message stating that the current section was not found to	The terms identified in the worldcloud are not necessarily helpful. Personally, you should consider why someone would read the Privacy Policy. Terms like "Your data" or "your actions" are key as this is what i am interested in. So, i would like to have this highlighted in the text so i can be draw to them.	yes - another section that will give one to two sentences bullet point of the policy	
It took a while to find the 'general information' so i could read on how to process the information the site was giving. Also the explanations given in 'general information' about each of the four tabs could be also presented when viewing the tab.	would prefer the font to remain the same throughout the page (referring to the title and "insert policy" instructions). Liked the dark background because it's easier for the eyes, but some people may prefer otherwise when it comes to important docs.		Add more languages or open the English site by default. Manually mark areas that contain important information but were not detected by the application so it can be improved. Add more specific categories in the pieChart instead of "other".	
Many privacy policies don't end in .html and therefore didn't work.	The text part is useful and the filters help but it could be friendlier. Improvements on text formatting would help - so maybe keep original formatting (titles, bullets etc) would be better.		You should suggest tags in the text in order to find more info. You could suggest a specification to provide a beautifier in websites or in browsers.	
In the pie chart presentation categories other and international should have more dissimilar colours,	Policy tab: Don't show options for parts that not existed.		In the textual form, when i select a certain category, I would prefer to load only the text	
	Maybe on hover of the tags have a clear description of what that filter actually filters		Pie chart, wordcloud and GDPR should have hyperlinks to relevant sections in the Policy tab.	
	It would be nice to distinguish between the selected filter for the textual representation in a different manner. Suggestion: Provide focus to the filtered text rather than removing focus from the other filters (by using a small font size).		the links on the left don't necessarily work as expected. Would expect them to either pop up the respective information, or at least scroll down to the relevant section.	
	less text right on the subject/ target		capability of applying multiple filters on Textual presentation	
	The classification buttons should be disabled when i clicked it for the second time, instead of scrolling for the clear button		Include word search.	
	Makenit apply to more urls			
	In the pie chart presentation categories other and international should have more dissimilar colours, because its hard to distinguish them.			

Figure 5.3.2.13 Changes, additions, complaints and requests as given by the users

Chapter 6

Conclusion, Discussion and Future Work

6.1 Introduction	67
6.2 Conclusion	67
6.3 Discussion	69
6.4 Future Work	70

6.1 Introduction

During the research, development and analysis that took part for the creation of this thesis project, some conclusions have been reached. These conclusions will be analysed below along with a discussion about the project. Finally, future work that can be done to improve this project as well as other projects or research that can use this work as a stepping stone will be mentioned and analysed.

6.2 Conclusion

In the context of this thesis a web application was designed, created and deployed to help end users increase their privacy awareness. This web application named Privacy Policy Beautifier is designed to take the URL or text of a privacy policy given by the user and present that privacy policy in a more user friendly way to help the users find what they are interested in.

This transformation is done with the help of a classifier that was specifically trained to classify segments of text into classes based on the topic they are referring to. These classes are taken from the OPP-115 dataset that is provided

by [31], as well as the data used for training and testing the classifier. The classifier that was created and used has an accuracy of 74% when it comes to classifying text that was not used during its training phase. To achieve this level of accuracy the training data was processed in various different ways, this included removing stop words that provided no useful information, lemmatizing and vectorising words. Additionally, different classifiers were tested, each with their own advantages and disadvantages. These classifiers included: Multilayer Perceptron (MLP), Naive Bayes and the Random Forest classifier, which is the one used in the latest version of the Privacy Policy Beautifier.

Afterwards, the now classified privacy policy is sent to the web application, which in turn presents it to the user as beautifully and clearly as possible. This is done in several ways, like the use of colours to indicate the different categories contained in the privacy policy, with a pie chart that summarizes the contents of the text, or a word cloud that presents the most frequently appearing words. Moreover, a two dimensional table showing the inclusion of GDPR terms in the privacy policy is also available for the user. The decision to use colors, pictures and 2D tables comes from the findings of papers like [26], which shows that users find them more appealing and easy to read.

While creating the classifier some limitations were encountered. The main issue was the limited amount of annotated privacy policies that exist that can be used for training the classifier. Fortunately, there is a plethora of techniques that can be used for pre-processing data and a wide range of classifiers, each with their own set of parameters to be modified. Every one of these may have a positive or negative effect on the accuracy of the system but so can the huge numbers of combinations between them. This large number of possibilities brings hope that with more time and testing, the accuracy of the system can be improved in future versions. Furthermore, unsupervised techniques as seen in papers like [24] or a combination of supervised and unsupervised techniques can be used as shown in papers like [10].

6.3 Discussion

The Privacy Policy Beautifier was created in order to help users raise their privacy awareness, which is relatively low. This is an issue that has been observed by papers like [22, 26]. Several studies, as well as this one, have found that this issue is caused by the length of privacy policies, the amount of time it takes to read them and the vague and confusing language they use [2, 17, 19, 21]. Fortunately, other studies, including this one, have found that users respond positively to attempts made to improve the visualization of privacy policies [14,18,26].

Various attempts have been made by researchers and other organizations to minimize this issue. Governing bodies like the European Union and the United Nations have approved and enforced legislations and regulations to protect the right to privacy of every human [5, 9, 23]. Whereas, researchers have created tools and systems to help users be more informed and knowledgeable when it comes to knowing how and where their personal information is being used [1,10,13,24,32].

All these attempts take a different approach to helping users, whether that is to the visual or the technical aspect. The Privacy Policy Beautifier uses information from these studies to try and improve upon them. This information includes things like the use of colors, 2D tables and pictures, as well as ways to train the classifier in the backend, which is the part doing most of the heavy lifting in this process. The Privacy Policy Beautifier uses supervised learning to train the classifier, whereas papers like [24] use unsupervised learning and tools, like Pribot [10], that uses a combination of the two methods, all with varying results.

The Privacy Policy Beautifier is mainly meant to be used by everyday users to help them find the information they are looking for from the wall of text that is a privacy policy. This does not mean that companies and organisations can't benefit from using this web application. Companies can use the Privacy Policy

Beautifier to see how their privacy policy will be seen by users or to think of ways to make the original privacy policy more user friendly to begin with. Additionally, smaller businesses that have trouble being compliant with regulations like the GDPR [6] can use aspects of the Privacy Policy Beautifier to help them out.

6.4 Future Work

During the creation of the Privacy Policy Beautifier some hurdles had to be overcome. Mainly, the lack of annotated privacy policy datasets to help train the classifier. In the future, more focus should be given to improving the accuracy of the classifier. This can be done by finding more data for training, by trying different combinations of techniques or algorithms, or by having a more efficient pre-processing step for the raw data. Additionally, a database could be used to track the most common privacy policies used in the web application. Moreover, the Privacy Policy Beautifier can be improved by allowing users to insert any URL, not just ones ending in “.html”, or even implement a crawler that finds the privacy policy page of a site automatically, so the user will only have to insert the URL of the site's main page. Moreover, the “GDPR term” tab could be expanded to and improved to maybe include a score or a percentage of GDPR coverage. Furthermore, more options of different visualisation can be added containing information for other aspects of the privacy policy that users might find useful or interesting. Also, a database with the most used or searched privacy policies can be created to gather information on what users are most interested in seeing as well as finding ways to improve the web application even further

Finally, any other study focusing on the readability of privacy policies or the effect it has on users, can use the findings of this study to aid them. Additionally, the Privacy Policy Beautifier can be used in combination with other questionnaires to try and probe deeper into how users experience variations of privacy policies as well as their willingness to spend the necessary time to read what is important to them.

Bibliography

- [1] Ζαμπά, Γιώργος. “Evaluating privacy policies on web platforms based on the GDPR.” 2020.
- [2] Aïmeur, Author links open overlay panelEsma, et al. “When changing the look of privacy policies affects user trust: An experimental study.” *Computers in Human Behavior*, vol. 58, 2016, pp. 368-379. *ScienceDirect*, <https://www.sciencedirect.com/science/article/abs/pii/S0747563215302296>.
- [3] Angulo, Julio, et al. “Towards usable privacy policy display and management.” *Information Management & Computer Security*, vol. 20, no. 1, 2012. *Emerald logo*, <https://www.emerald.com/insight/content/doi/10.1108/09685221211219155/full/html>.
- [4] B.H.M, Custers. “Predicting Data that People Refuse to Disclose; How Data Mining Predictions Challenge Informational Self-Determination.” *Privacy Observatory Magazine*, vol. 2012, no. 3. *Leiden University Repository*, <https://openaccess.leidenuniv.nl/handle/1887/46935>.
- [5] Council of Europe. “The Convention for the Protection of Human Rights and Fundamental Freedoms.” 1950. *council of Europe*, <https://www.coe.int/en/web/human-rights-convention/the-convention-in-1950#:~:text=The%20Convention%20for%20the%20Protection,for ce%20on%203%20September%201953>.

- [6] Espinoza, Javier. "EU admits it has been hard to implement GDPR." *Financial Times*, 23 June 2020, <https://www.ft.com/content/66668ba9-706a-483d-b24a-18cfbca142bf>.
- [7] Galle, Matthias, et al. "The Case for a GDPR-specific Annotated Dataset of Privacy Policies." *CEUR Workshop Proceedings*, http://ceur-ws.org/Vol-2335/1st_PAL_paper_5.pdf.
- [8] "GDPR Enforcement Tracker." *GDPR Enforcement Tracker*, <https://www.enforcementtracker.com/>.
- [9] General Assembly resolution 2200A (XXI). "International Covenant on Civil and Political Rights." 1966. *United Nations Human Rights*, <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx>.
- [10] Harkous, Hamza, et al. "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning." *27th USENIX Security Symposium*, 2018. *27th USENIX Security Symposium*, <https://www.usenix.org/conference/usenixsecurity18/presentation/harkous>.
- [11] Hughes, Kirsty. "A Behavioural Understanding of Privacy and its Implications for Privacy Law." 2012. *wiley online library*, <https://doi.org/10.1111/j.1468-2230.2012.00925.x>.
- [12] Jeff, Smith H. "Managing privacy : information technology and corporate America." <https://archive.org/details/managingprivacyi0000smit>.

- [13] Kelley, Patrick Gage, et al. "A "Nutrition Label" for Privacy." *Proceedings of the 5th Symposium on Usable Privacy and Security*, 4:1–4:12, <https://cups.cs.cmu.edu/soups/2009/proceedings/a4-kelley.pdf>.

- [14] Kim, KyungTae, et al. "WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora." <http://users.umiacs.umd.edu/~elm/projects/wordbridge/wordbridge.pdf>.

- [15] Kumar, Vinayshekhar Bannihatti, et al. "Quantifying the Effect of In-Domain Distributed Word Representations: A Study of Privacy Policies." 2019. *USABLEPRIVACY.ORG*, https://usableprivacy.org/static/files/kumar_pal_2019.pdf.

- [16] Kupfer, Joseph. "Privacy, Autonomy, and Self-Concept." <https://philpapers.org/rec/KUPPAA>.

- [17] Lebanoff, Logan, and Fei Liu. "Automatic Detection of Vague Words and Sentences in Privacy Policies." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3508-3517. *ACL Anthology*, <http://dx.doi.org/10.18653/v1/D18-1387>.

- [18] Linden, Thomas, et al. "The Privacy Policy Landscape After the GDPR." *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, 2015, p. 64. *sciendo*, [https://content.sciendo.com/configurable/contentpage/journals\\$002fpopets\\$002f2020\\$002f1\\$002farticle-p47.xml](https://content.sciendo.com/configurable/contentpage/journals$002fpopets$002f2020$002f1$002farticle-p47.xml).

- [19] Liu, Fei, et al. "Modeling Language Vagueness in Privacy Policies using Deep Neural Networks." *Cornell University*, <https://arxiv.org/abs/1805.10393>.

- [20] Lomas, Natasha. "GDPR's two-year review flags lack of 'vigorous' enforcement." 24 June 2020, https://techcrunch.com/2020/06/24/gdprs-two-year-review-flags-lack-of-vigorous-enforcement/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAACZsbHFen3IV8l68DjCH1D9-ciY2bp8MV_jWLSauFp4ln2gPIVQfMJXTPWn1DwlQXaD9fKJY_CH2uCmJLqV.

- [21] McDonald, Aleecia M., and Lorrie Faith Cranor. "The Cost of Reading Privacy Policies." https://kb.osu.edu/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf.

- [22] Pitkänen, Olli, and Virpi Kristiina Tuunainen. "Disclosing Personal Data Socially — An Empirical Study on Facebook Users' Privacy Awareness." *Journal of Information Privacy and Security*, vol. 8, no. 1, 2012. *Research Gate*, 10.1080/15536548.2012.11082759.

- [23] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. "REGULATIONS." *Official Journal of the European Union*, 2016. *GENERAL DATA PROTECTION REGULATION (GDPR)*, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.

- [24] Sarne, David, et al. "Unsupervised Topic Extraction from Privacy Policies." *WWW '19: Companion Proceedings of The 2019 World*

Wide Web Conference, 2019, pp. 563-568. *ACM DIGITAL LIBRARY*, <https://dl.acm.org/doi/abs/10.1145/3308560.3317585>.

- [25] Solove, Daniel J. "Understanding Privacy." *Harvard University Press*, 2008. *SSRN*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1127888.
- [26] Soumelidou, Aikaterini, and Aggeliki Tsohou. "Effects of privacy policy visualization on users' information privacy awareness level: The case of Instagram." *Information Technology & People*, vol. 33, no. 2, 2019. *emerald insight*, <https://www.emerald.com/insight/content/doi/10.1108/ITP-08-2017-0241/full/html>.
- [27] Strobel, Hendrik, et al. "Document Cards: A Top Trumps Visualization for Documents." *IEEE Transactions on Visualization and Computer Graphics*, 2009, <https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1109%2FTVCG.2009.139>.
- [28] United Nations General Assembly. "Universal Declaration of Human Rights." 1948. *United Nations*, <https://www.un.org/en/universal-declaration-human-rights/>.
- [29] Warren, Samuel D., and Louis D. Brandeis. "The Right to Privacy." *Harvard Law Review*, vol. 4, no. 5, 1890, pp. 193-220. *Cornell CIS Computer Science*, <https://www.cs.cornell.edu/~shmat/courses/cs5436/warren-brandeis.pdf>.

- [30] Westin, Alan F. "Privacy And Freedom." *Washington and Lee Law Review*, vol. 25, no. 1, 1968. *Washington and Lee University school of law*, <https://scholarlycommons.law.wlu.edu/wlulr/vol25/iss1/20/>.
- [31] Wilson, Shomir, et al. "The Creation and Analysis of a Website Privacy Policy Corpus." 2016. *USABLEPRIVACY.org*, https://usableprivacy.org/static/files/swilson_acl_2016.pdf.
- [32] Zaeem, Razieh Nokhbeh, et al. "PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining." *ACM Transactions on Internet Technology*, 2018. *ACM DIGITAL LIBRARY*, <https://dl.acm.org/doi/abs/10.1145/3127519>.

Appendix A

Questionnaire

Privacy Policy Beautifier

The following questions have to do with your experience reading and understanding privacy policies of any application.

Informed Consent Form

This survey does not have any commercial purposes, the involved researchers do not have any monetary benefits by conducting it and the results will be published in the form of reports and research papers based on the survey. This questionnaire is anonymous. You will not be asked to provide any information that may reveal who you are or that may be traced back to you.

By responding to this questionnaire, you confirm the following:

- I have read and understood the purpose of the survey.
- I understand that my taking part is voluntary. I can withdraw from the study at any time during the survey and I do not have to give any reasons for why I no longer want to take part.
- I agree that the answers I give will be stored in digital form. Only the involved researchers will have access to this information and this information will not be distributed to another person or entity.

For more information, please contact the involved researchers:

Michalis Kaili - mkaili02@cs.ucy.ac.cy

Georgia Kapitsaki (supervisor) - gkapi@cs.ucy.ac.cy

***Required**

*

☐ I provide my consent to the above

Privacy Policy Beautifier

*Required

General Information

What is your age? *

- ☐ Under 18
- ☐ 18 - 24
- ☐ 25 - 29
- ☐ 30 - 39
- ☐ 40 - 49
- ☐ 50 - 60
- ☐ Above 60

Gender *

- ☐ Male
- ☐ Female
- ☐ Prefer not to say

Would you say you are a technical expert? *

- ☐ Yes
- ☐ No
- ☐ I know some stuff

What level of education have you completed or are you currently attending? *

- ☐ Middle School
- ☐ High School
- ☐ Bachelor
- ☐ Master
- ☐ PhD
- ☐ Other: _____

Educational Background

What are you studying/have you studied?

Your answer _____

Privacy Policy Experience

Have you ever read a privacy policy from any website or application? *

- ☐ Yes
- ☐ No
- ☐ Partially

Read a privacy policy

Was your reading experience enjoyable? *

- | | | | | | | |
|-------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------|
| | 1 | 2 | 3 | 4 | 5 | |
| Strongly Disagree | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Strongly Agree |

Did you find the information you were looking for easily? *

- | | | | | | | |
|-------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------|
| | 1 | 2 | 3 | 4 | 5 | |
| Strongly Disagree | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Strongly Agree |

Would you ever read another privacy policy? *

- ☐ Yes
- ☐ No
- ☐ Maybe

Would not read another privacy policy

Why would you not read another privacy policy?

Your answer _____

Might read another privacy policy

What would make you read a privacy policy?

Your answer _____

What would make you not want to read a privacy policy?

Your answer _____

Have not read a privacy policy

What reasons made you not want to read a privacy policy in the past? *

- ☐ I didn't care
- ☐ It needed too much time
- ☐ Hard to understand / Confusing
- ☐ Couldn't find it
- ☐ Too long
- ☐ I didn't need to
- ☐ Other: _____

Privacy policy Beautifier experience

The following questions have to do with your experience using the website privacy policy beautifier

Please use the following link to find our website and use it for a bit before answering the rest of the questions.

<http://privacypolicybeautifier.cs.ucy.ac.cy:5000/>

Feel free to use any of the following links to privacy policies in the Privacy Policy Beautifier:

Make sure you insert the entire URL as seen below or click the link and copy it from the site itself

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

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

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

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

Did you find it easy to use the privacy policy beautifier website? *

	1	2	3	4	5	
Very Hard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Easy

Was the privacy policy easier to read? *

	1	2	3	4	5	
Strongly Disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly Agree

Would you consider using privacy policy beautifier again? *

- ☐ Yes
- ☐ No
- ☐ Maybe

Which presentation/s of the content of the privacy policy did you prefer? *

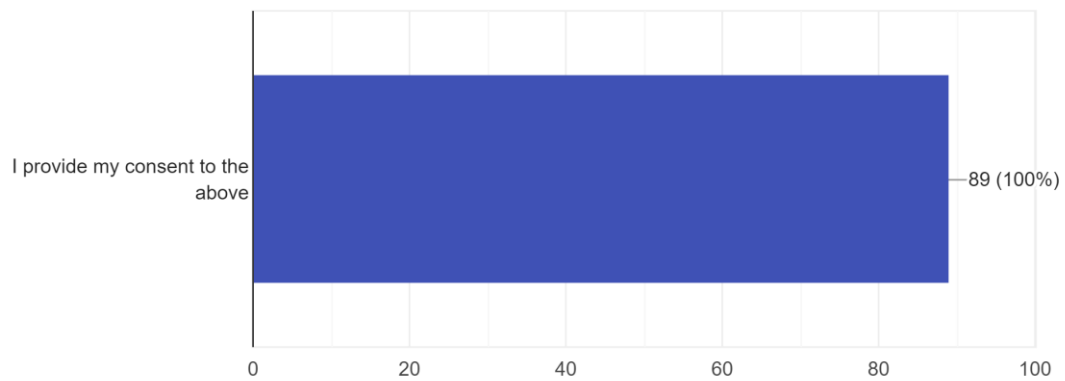
- ☐ Textual
- ☐ PieChart
- ☐ WordCloud
- ☐ GDPR (General Data Protection Regulation) keys

Do you have any comments/suggestions about the site how it is presented and how you think it could be improved?

Your answer

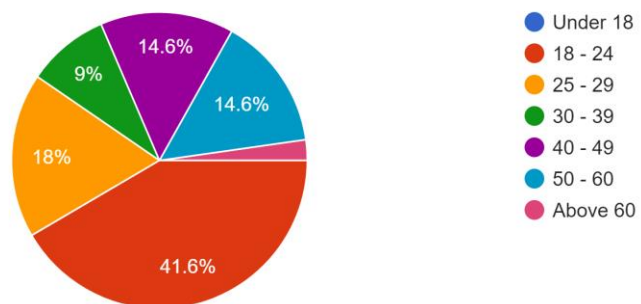
Questionnaire Answers

89 responses

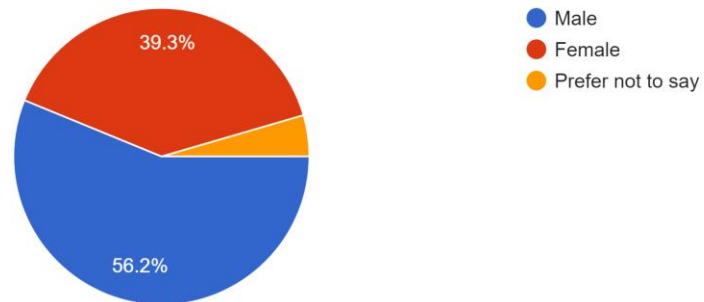


What is your age?

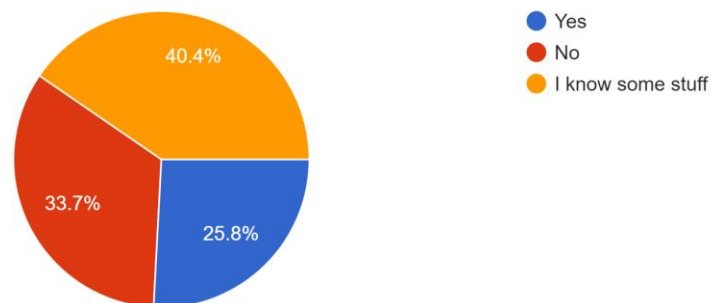
89 responses



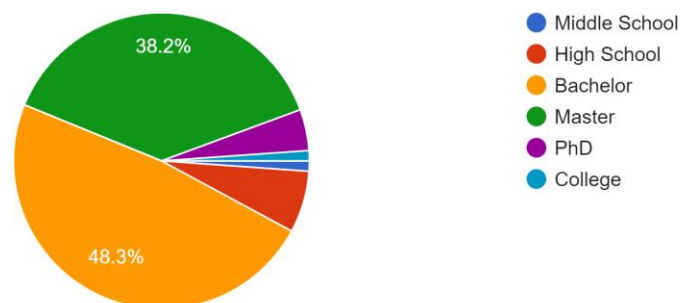
Gender
89 responses



Would you say you are a technical expert?
89 responses



What level of education have you completed or are you currently attending?
89 responses



What are you studying/have you studied?

81 responses

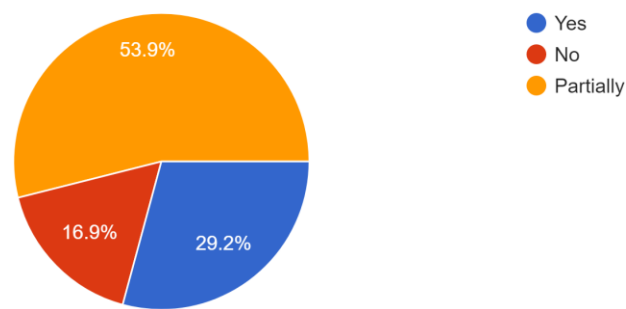
	Frequency
	8
Accounting and Finance	1
agriculture	1
BA Studio Art, Art History, MA Arts Management and Administration	1
banking, Physical education	1
Bioinformatics	1
Biological sciences	1
Biology	1
Bsc Computing	1
Business	1
Business Administration	1
Chemical engineering	1
Chemistry	1
Civil Engineer	1
Civil Engineering	2
Computer Engineering	2
Computer Science	20

Computer Science/Web and smart Systems	1
Criminology	1
Culinary arts	1
Data Communication Systems	1
Economics	1
Education	1
electrical and computer engineering	1
Electrical and electronics engineering	1
Electrical Eng and MBA	1
Electrical Engineering	6
Electrical Engineering / Computers	1
Electrical&Computer Engineering/Telecoms	1
Engineering	1
Engineering/ Management	1
English language and literature	1
English literature and Psychology	1
English Studies	1

Environmental Engineering	1
Finance	1
Human Resources	1
I have studied	1
international relations	1
Journalism	1
Law	1
Marketing	1
Mathematics	2
Mathematics and statistics	1
maths	1
Mechanical engineering	1
Medicine	1
mining engineering	1
MProf strategic business support	1
MSc Interaction Technology	1
Occupational Therapy	1
physiotherapy	1
Psychology	2
Studies Mass Media and Marketing	1
Total	89

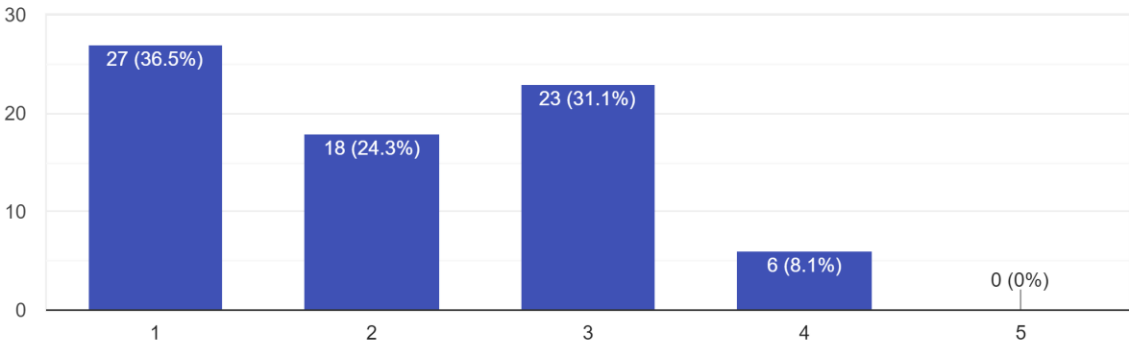
Have you ever read a privacy policy from any website or application?

89 responses



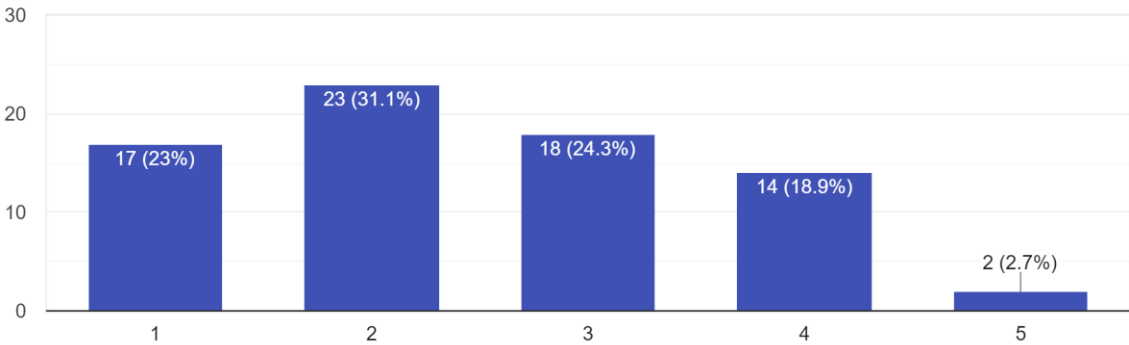
Was your reading experience enjoyable?

74 responses



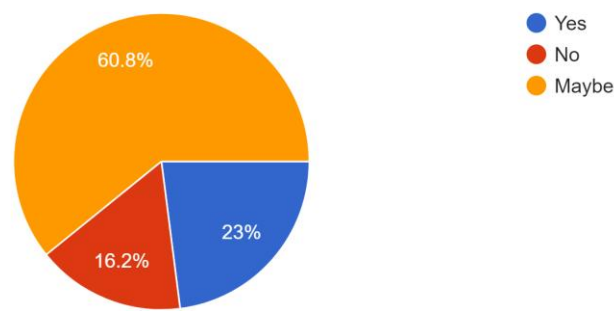
Did you find the information you were looking for easily?

74 responses



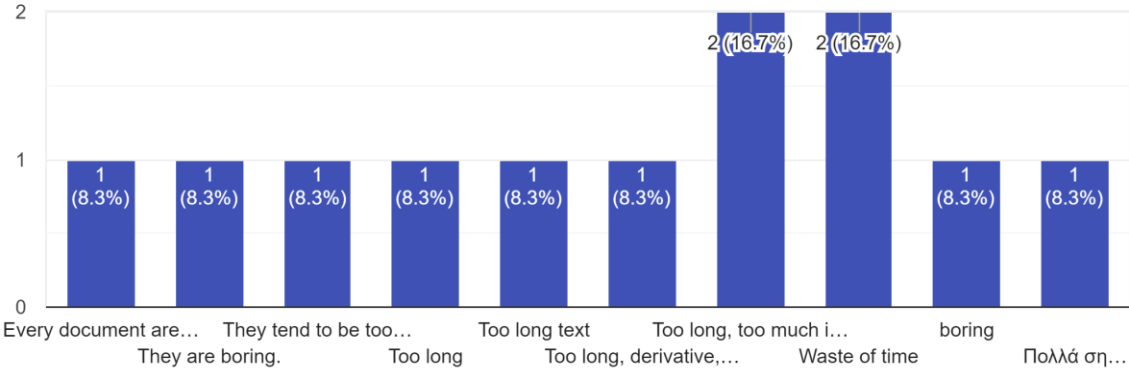
Would you ever read another privacy policy?

74 responses



Why would you not read another privacy policy?

12 responses



What would make you read a privacy policy?

80 responses

	Frequency
	9
a simple and compact PP	1
A brief document with clear, bullet form points outlining the policy.	1
A summary of what I should really care. What is the worst case scenario that can happen. How is it different from competitive policies.	1

Absolute need to access the website	1
Any policies related to Mark Zakenberg	1
Anything that makes user ID known to others	1
Big red letters urging me to read it	1
bullet point summary	2
Bullet points	2
Check data access and sharing	1
Clearly formatted text with links to each section and searching functionality	1
Concerns about my privacy	1
Copyrights	1
curiocity	1
curiosity	1
Depending on the site/app/permissions requested	1
Depends on the severity of the application	1
Details	1

easy read with segments	1
For privacy and security issues	1
Had to	1
I had to write my own for some websites that I have.	1
i want that all policies are clear	1
I would read it anyway	1
If I am looking for something specific	1
If I share a lot of sensitive information (mostly bank accounts and such) I would like to know how they are being handled.	1
If I thought my privacy would be severely compromised if I didn't read it but accepted anyway	1
If is a short paragraph	1
If it has to do with my money	1
If it is short & concise	1
If it was more concise	1
If it was requires from me	1
If it was short	1

If it was short and to the point	1
If it was shorter & easier to understand	1
If it was shorter.	1
If its two linea	1
If the matter affects me significantly	1
Important	1
Know the Data the Privacy policy will collect. Simple Policy	1
legal concerns	1
Make it like a sectioned form with next and back buttons so that it's easier to traverse, have it written as simply as possible	1
Make it more concise	1
money	1
My privacy safety	1
need for information	1
Not mainstream/common app/website/form	1
nothing	1

Only bank's terms and conditions. They are liars...	1
Only few bullets	1
Only if I had a specific interest (e.x if my personal data are used in any way)	1
only the important	1
Protection of personal data	1
Safety reasons	1
Searching for specific rules the company/website is applying	1
short and concise text	1
short and understandable material	1
Short and with points of interest	1
simple presentation	1
Small clear document	1
small short bullet points divided in topics	1
Small Size of document	1

Smaller size	1
Something with smaller text	1
the approach of the company/sender regarding the relevant issue	1
The fear that my data are used for purposes that I am not aware.	1
The user interface	1
to be short and simple	1
To find what I am looking for	1
to make sure of my legal rights	1
To see how my data is being used, especially when I plan to use said service often	1
Trust issues	2
trust issues, know how my data will be used	1

Understand the privacy policy of a website/application that I might be or intend to register at, in order to know their policies around data handling and security.	1
User friendly - necessary info	1
When i will develop a new application and i will use a software from other person and i will want to see the licence	1
Την πρώτη φορά δεν ήξερα τι ήταν, οπότε την διάβασα	1
Total	89

What would make you not want to read a privacy policy?

85 responses

	Frequency
	4
3 pages of writings	2
a long multi page complicated PP	1
A long one	1

A long one with many words that I do not understand	1
A long policy	1
a lot information	1
A lot of blabla	1
A lot of legal terms that I do not understand	1
A lot of stuff	1
A lot of text	1
Anything else	1
Big texts	1
boredom	1
Boredom	1
Complicate Policy with more that 1 page	1
Huge document	1
I know what they tell.	1
If I am not looking for something specific and don't have the enough time	1

If is a lot to read	1
If it is too long	1
If it is too long or not worth reading	1
If it looks good	1
if it looks too large and overwhelming	1
If it was long	1
If it was long with lots of jargon that was hard to understand	1
If it was too long	1
If it was too long, and contained too many technical terms	1
If it was very long to read.	1
If it's very long	1
If the length is too long	1
If the subject was unimportant	1
It's length	1
Its extent	1
Its length	1

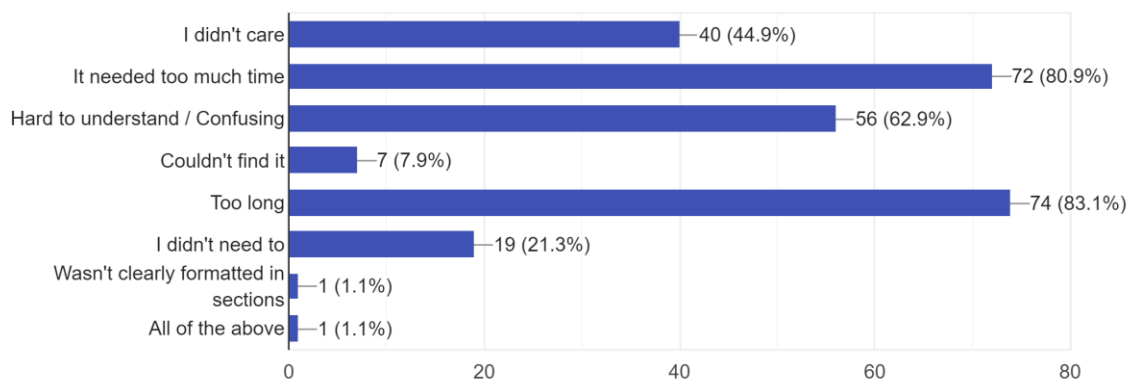
Its time consuming	1
its too long	1
Knowing that it doesn't contain what I am looking for	1
Large amount of unnecessary content	1
lenght	1
length	1
Length, appearance, things I do not need to know	1
long	1
Long and incomprehensive text	1
Long article	1
Long Complicated stories written...	1
Long document	1
Long plain text	1
long read	1
long sentences with no specific order	1
long text	1
Long text	1
Long text without formatting and no links to each point,	1
Long text, complex vocabulary	1

Long texts	1
Long winded vocabulary	1
need for swiping, elaborate legal statements	1
Ridiculously long text/ legal language	1
so many information, people want to continue doing their stuff	1
takes too much time	1
the complex language and long text	1
The fact that is way too long.	1
the size of the document	1
Time	1
Time limit/ In a hurry and consider it irrelevant based on task	1
Tiny letters with bad font and too many words	1
To many details, it looks long at first glance	1
Too big context	1

Too complicated legal terms / very long policy	1
Too long	4
Too much information	1
Too much information without a good user interface	1
Too much information-long text	1
too much information/ no available time	1
Too much reading and difficult to understand	1
Too much text to read	1
Trust in the company.	1
Updates of mainstream app/products etc	1
What I am accountable for.	1
When i install one video game	1
When its applied in video games	1
When the corresponding site or company will not own of the previous mentioned sensitive information.	1
Total	89

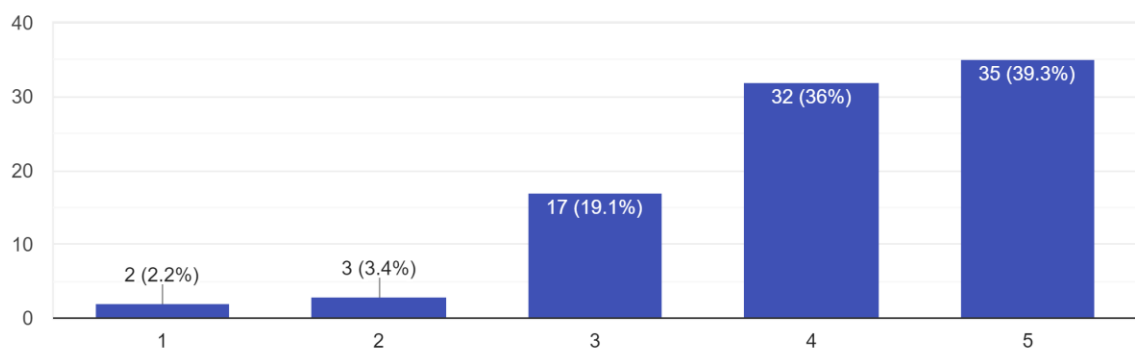
What reasons made you not want to read a privacy policy in the past?

89 responses



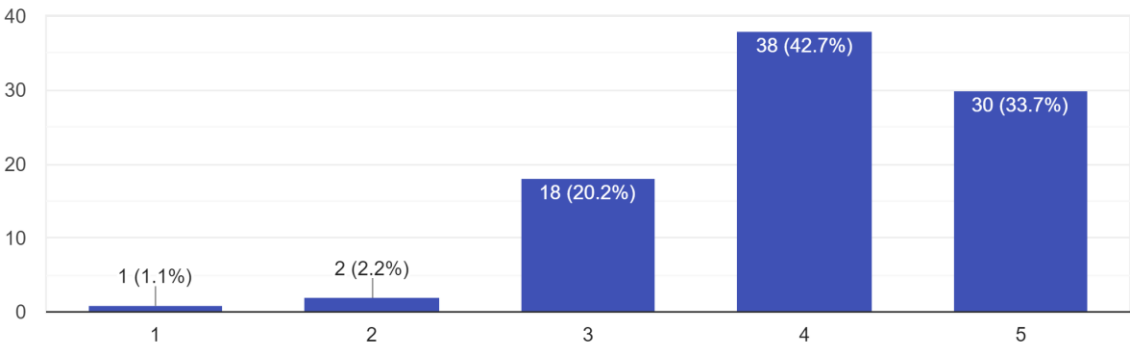
Did you find it easy to use the privacy policy beautifier website?

89 responses



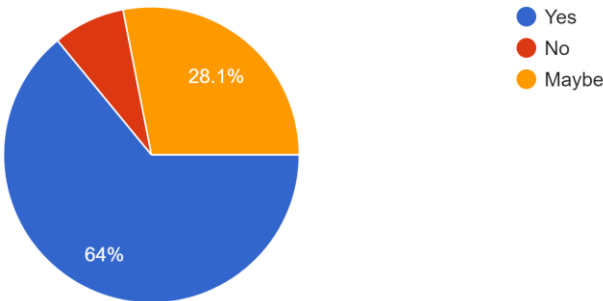
Was the privacy policy easier to read?

89 responses



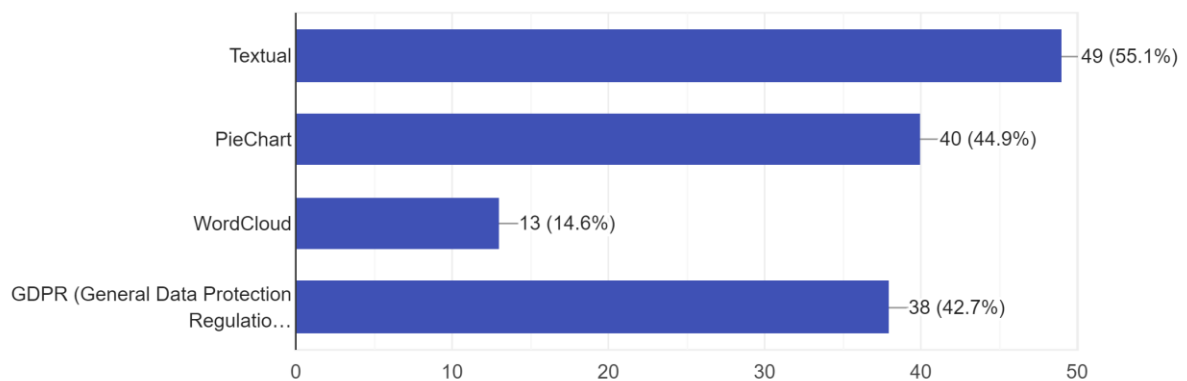
Would you consider using privacy policy beautifier again?

89 responses



Which presentation/s of the content of the privacy policy did you prefer?

89 responses



Do you have any comments/suggestions about the site how it is presented and how you think it could be improved?

44 responses

	Frequency
	45
1. The word cloud is not useful	1
Add more languages or open the English site by default. Manually mark areas that contain important information but were not detected by the application so it can be improved. Add more specific categories in the pieChart instead of "other".	1

after pressing button it would be nice to be transferred automatically to that highlighted section. it would be nice if the categories that are not covered in the privacy policy were put in a separate section so that the user does not click on button that do nothing OR if the button does nothing display a message stating that the current section was not found to be covered in the privacy policy. overall I am very impressed with this tool and it is something i would definitely use!	1
capability of applying multiple filters on Textual presentation	1

Difficult to find privacy policy links strictly ending in .html. Presentation wise: would prefer the font to remain the same throughout the page (referring to the title and "insert policy" instructions). Liked the dark background because it's easier for the eyes, but some people may prefer otherwise when it comes to important docs.	1
Easy to answer and to understand	1
Explain to people that even if they copy an address that does not end with html, all they need to do for the beautifier to work is to type an html at the end.	1

I believe it is a great tool that can help users finding the information they need	1
If it could give me a summary of the section of interest	1
In the pie chart presentation categories other and international should have more dissimilar colours, because its hard to distinguish them. I don't find the wordcloud useful, maybe if it was per category it would be more meaningful.	1
In the wordcloud i think different size and colours of the words-appeared may confuse the reader and make difficult to catch the point	1

Include word search.	1
Is it perfect. Any change.	1
It looks like an error page	1
It made understanding policies much simpler.	1
It took a while to find the 'general information' so I could read on how to process the information the site was giving. Also the explanations given in 'general information' about each of the four tabs could be also presented when viewing the tab.	1
It would be nice to distinguish between the selected filter for the textual representation in a different manner. Suggestion: Provide focus to the filtered text rather than removing focus from the other filters (by using a small font size).	1

Its perfect	1
less text right on the subject/ target	1
Makenit apply to more urls	1
Many privacy policies don't end in .html and therefore didn't work.	1
Maybe on hover of the tags have a clear description of what that filter actually filters	1
no	2
No	2
No I think it was pretty straight foreword	1
No it is very good !	1
No, everything was perfect. :)	1
overall a nice site, never thought of the alternative ways to present privacy policy	1

Policy tab: Don't show options for parts that not existed.	1
Privacy policies will always be boring to read	1
Set the document into pieces/categories	1
Small explanations about the different things in pie chart and textual window for the people that have no clue about the topic	1
some colors just be easier to read at headlines	1
The classification buttons should be disabled when i clicked it for the second time, instead of scrolling for the clear button	1

The text part is useful and the filters help but it could be friendlier. Improvements on text formatting would help - so maybe keep original formatting (titles, bullets etc) would be better.	1
Very nice	1
Very nice presentation and fast response	1
Was pretty good	1
yes - another section that will give one to two sentences bullet point of the policy	2
Yes.	1
You should suggest tags in the text in order to find more info. You could suggest a specification to provide a beautifier in websites or in browsers.	1
Total	89

