



Data practices in apps from Brazil: What do privacy policies inform us about?

Valéria Quadros dos Reis  [Federal University of Mato Grosso do Sul | valeria.reis@ufms.br]

Maria E. R. Rabello  [Federal University of Mato Grosso do Sul | elisa.rabello@ufms.br]

Anderson C. Lima  [Federal University of Mato Grosso do Sul | anderson.lima@ufms.br]

Guilherme P. S. Jardim  [University of Campinas | guilhermepsjardim@gmail.com]

Eraldo R. Fernandes  [Universität Leuphana | eraldo.fernandes@leuphana.de]

Ulf Brefeld  [Universität Leuphana | brefeld@leuphana.de]

Abstract

Abstract We present and analyze a new corpus comprising 82 privacy policies in Brazilian Portuguese collected from popular apps in Google Play Store called APP-BR. The contained documents are characterized by excessive lengths and poor readability. Analyzing their content reveals a severe lack of objectivity and compliance with the Brazilian General Data Protection Law. Our results shed light on the problems in accessing privacy information and are supposed to constitute a basis for finding a remedy. To the best of our knowledge, this is the first effort to evaluate privacy policies according to the current Brazilian legislation.

Keywords: *privacy corpus, mobile application, LGPD, personal data, law compliance*

1 Introduction

The Brazilian General Data Protection Law, known as LGPD, regulates the collection and use of personal data (L13709, 2018). According to this law, digital services, such as mobile apps and websites, must ask users for their consent before processing their data. Due to this mandatory agreement between organizations and individuals, privacy policy contracts became popular sources of information (Siebra and Xavier, 2020).

Nonetheless, many app providers neglect users access to their privacy policies. Abreu (2018) analyze 13 apps developed by the Brazilian federal government and observe that at least 6 of them do not state their privacy policies. Moreover, many companies only provide incomplete privacy policies that do not disclose the full extend of data processed. Barbosa (2017) reveal that some apps in the Google Play Store, missed to update the corresponding privacy policies even though an update of the app changed the data access permissions. The authors also analyze source codes and found severe inconsistencies between data processing implementations and privacy policies.

Another important issue comprises privacy policies that are difficult to understand. There are multiple reasons for this, including overly long texts, unappropriated language, and ambiguity in phrasing. Pollach (2007) analyzed 50 privacy policies from North American websites and conclude that such documents need more user-centered content and user friendly presentations. Still considering the North American scenario, Fabian et al. (2017) measure the readability of privacy policies for almost 50,000 websites. Results show that, on average, privacy policies are characterized by poor readability, requiring a reader to have an expertise of between 9 to 15 years to be able to fully grasp the content. The mentioned problems in privacy policies are frequently pointed out as reasons why people do not read them before using digital services (Obar and Oeldorf-Hirsch, 2018).

To the best of our knowledge, our previous work (Jardim et al., 2022) was the first effort to characterize privacy policies from Brazil through the creation of a corpus. We shed light on the access patterns of 1,163 policies crawled from apps in Brazil and observe that only 10% of these documents are original and actually written in Portuguese. We also report on broken links and overly complex texts that unnecessarily bedevil understanding.

In the present paper, we continue to profile privacy texts of the previously published corpus, but rather focus on semantics. To do so, we manually label every paragraph of 82 privacy policies according to data practices proposed and validated in Wilson et al. (2016). Our hypothesis is that this annotation process may reveal the main topics covered by privacy policies and highlight the level of law compliance of these documents when considering the LGPD. Our tagged corpus, denoted *APP-BR*, is also relevant for the implementation of services that automatically extract information from privacy texts. Examples include, but are not limited to, text summarization and law enforcement audit.

The remainder is organized as follows. Section 2 presents important concepts used throughout the text and Section 2 reports on related work. Sections 4 and 5 discuss the creation and analysis of the corpus, respectively. In Section 6, we analyze the compliance of the policies to the LGPD. Section 7 highlights threats to the validity of our investigation and Section 8 considers the sociological aspects of our findings. Section 9 concludes.

2 Theoretical Aspects

2.1 Privacy Policies and Terms of Use

A privacy policy (PP) is usually presented as a contract that regulates the use of personal data by an organization providing a service or a product (Siebra and Xavier, 2020). A PP shall describe different aspects related to personal informa-

tion, such as which data is collected, how it is collected, for what purposes, how this data is secured, what is the period of retention, among others. A PP still needs to state whether the collected data is shared with third-parties and what is the procedure in case of policy changes.

Frequently, PPs are confused with terms of use. Terms of use, however, establish usage rules for services and state limits to the obligations of the organization (Yamauchi et al., 2016). In this work, we focus on privacy policies.

2.2 Legal Design

Legal Design is the application of human-centered design to legal documents, aiming at making them more usable and satisfying for users (Berger-Walliser et al., 2017). Examples of good practices in Legal Design are the presentation of short texts, the use of easy-to-understand terms, the creation of frequently asked questions, and the exposure of contractual clauses through the use of images. One common bad practice in privacy policies is the inclusion of advertisements within the text. Likewise, captchas and cookies are also undesired; the former render access to information difficult while the latter require access to user information even before the PP is available to the user.

2.3 Readability

Readability is a measure of how easy a text is to read. It depends on the characteristics of the document, such as presentation aspects (fonts, line height, character spacing, line length) and content aspects (complexity of vocabulary and syntax, for instance). Readability is also dependent on characteristics of the reader like literacy and domain familiarity. A legal document, for example, is easier comprehensible for experts than untrained people on all educational levels (Barboza and Nunes, 2008).

The Flesch Reading Ease is a well-established metric used to access the readability of texts in terms of their content (Flesch, 1979). Its adaptation for the Portuguese language is given by the following equation:

$$READ = 248,835 - (1,015 * ASL) - (84,6 * AWL),$$

where *ASL* is the average sentence length (number of words divided by the number of sentences) and *AWL* is the average word length (number of syllables divided by the number of words) (Martins et al., 1996). Flesch scores should vary between 0 and 100 but, eventually, when text patterns diverge from the average expected, they can extrapolate such values. The higher its value, the higher the readability, i.e., the easier to read the text by an arbitrary reader.

3 Related Work

Some related work stand out in the international scenario. Obar and Oeldorf-Hirsch (2018) conduct experiments with 543 attendants to present reasons why individuals ignore privacy policies. Among them are the excessive lengths of texts, the feeling that personal data does not need to be private, the need to use digital services at any price and the difficulty to

understand the documents. Our work corroborates the former and the last statement.

Fabian et al. (2017) reason about privacy policies problems. Their paper details the implementation of a privacy policy extraction and analysis tool. The authors obtain the privacy policies of more than 202,144 highly popular websites. Of these, only 163,232 pages had English content and, according to a ranking algorithm, only a third of the pages actually featured a PP. The analysis of the 49,036 proper policies reveal that the documents had an average size of 1,700 words and were difficult to understand, so that readers need higher levels of education than an American high school degree to comprehend the PPs. Another important contribution of this work is to present the strong correlation of the Flesch score with other scores widely used in the literature. The authors present a methodology and results similar to those performed in this work. Their conclusions also validated the readability score we used.

In 2016, (Wilson et al., 2016) release the annotated corpus OPP-115 consisting of 115 English privacy policies obtained from North American *websites*. Due to the robust methodology employed in its creation, OPP-115 represents an important step in assessing the quality of PPs. Such a valuable work inspired us to create the Brazilian analogue APP-BR and serves as a reference for our analysis.

In the Brazilian scenario, however, there is no publicly available privacy policy corpus. Nonetheless, a few works have systematically analyzed the content of PPs. Pontes (2016) use keywords and pattern matching algorithms to merge about 50 privacy policies collected from Brazilian websites into tabular structures. The authors state that structured presentations are easier understood by users, hence saving reading time and removing complicated legal terminology. Their classification was centered on the type of the collected user data, such as personal information and the last user activities. Our annotation is less detailed regarding user data but focuses on other aspects of data treatment, including user access to her information and data retention.

Viana et al. (2017) report results of a qualitative analysis on how terms of use and privacy policies address issues of postmortem digital legacy in five social web platforms. Although our research does not tackle specifically digital legacy after death, this feature is partially covered in our analysis under the practice about access, edit and deletion of user data.

Yamauchi et al. (2016) propose a set of ten guidelines to advise designers and developers in constructing understandable privacy interfaces. A contemporary work (Siebra and Xavier, 2020) presents a list of seventeen criteria to assess the quality and completeness of privacy policies. Although there is a great intersection among the two contributions and our analysis, all approaches raise distinct and important questions when evaluating PPs. For example, the present paper broadens the scope to also include childrens as users while Yamauchi et al. (2016) and Siebra and Xavier (2020), on the other hand, assess accessibility in general. Despite the fact that conformity to legislation is one criterion presented in the three works, to the best of our knowledge, our work is the first one to review policies considering the Brazilian LGPD.

4 APP-BR Corpus

4.1 Data Collection

The creation of the corpus has been performed in three phases: collection, preprocessing and analysis. Figure 1 presents some steps along these phases. Due to the large number of files to be analyzed and for ease of replication of experiments, these phases were largely automated.

To use a set of policies that complies with an indicator of social relevance, we collected policies from Google Play Store, one of the biggest stores of mobile apps in Brazil. The collection was assembled in February 2021 and resulted in 1,163 HTML pages, crawled from the privacy policy links of the most popular apps at that moment. We then manually selected only the documents in Portuguese. In this step, we identified many invalid files and documents that did not contain data policies. Such files, along with files that incompletely described data treatment, were removed from the initial group.

Finally, we de-duplicated the files and applied a boilerplate remover to eliminate HTML, CSS and JavaScript codes, as well as visible parts of the pages which did not contain PP text (menus, headers and footers, for instance).

4.2 Annotation Scheme and Process

We follow an annotation scheme similar to the one developed by Wilson et al. (2016) to create the OPP-115 corpus. We segment the texts into paragraphs and label each paragraph with one or more labels depending on the contained data practices. The set of labels was originally defined by an iterative refinement process, in which a group of domain experts (privacy experts, public policy experts, and legal scholars) identified different categories of data practice and their descriptive attributes from multiple privacy policies. In the original work by Wilson et al., each category is articulated by a category-specific set of attributes. However, for the sake of simplicity, we do not consider these associated attributes in our categories.

The data practices in OPP-115 are successfully validated in different experiments involving English texts (Wilson et al., 2016, 2018; d'Aquin et al., 2018). We decided to use this strategy as a reference to obtain a solid baseline for comparisons between PPs in English and other languages. Table 1 shows the categories of data practices used in the classification schemes.

Our annotation procedure involves three specialists in Computer Science that were previously trained on a small set of segments until they moderately agree among their classifications (Fleiss Kappa = 0.58). After the training phase, each collaborator was responsible for tagging an exclusive group of PPs. Subsequently, in order to improve consistency, all the annotations were reviewed by the most experienced annotator. The raw PPs and the annotated corpus are publicly available.¹ We make use of this corpus for quantitative and qualitative analysis in the remainder.

5 Results

5.1 Link Contents

Out of the 1,163 apps initially considered, only 926 presented valid links. Among the documents collected from these links, 715 (61.5%) of them were written in English and only 146 (12.6%) were actually written in Portuguese. The left side of Figure 2 presents the full results of the preliminary analysis. The right side of the figure shows a categorization of the files written in Portuguese. Among these files:

- 29 (20%) were discarded because they consisted of: terms of use (9), incomplete data practices (2), miscellaneous information (13) or another kind of non-privacy data. This reduced the number of files of interest to 117;
- 35 (24%) were discarded due to duplication. Big companies such as Meta, for instance, maintain the same policy for different apps (Whatsapp, Facebook and Instagram);
- Only 82 (56%) were distinct privacy policies.

Still considering the initial 146 Portuguese files, we highlight one good and one bad finding concerning legal design practices: agreement to the use of *cookies* were mandatory to access the contents of 26 files; on the other hand, a question-answering format was presented in 32 files.

We observe that, among a set with 1,163 documents, only 117 (10%) of them correspond to the information expected.

5.2 Document Length

Figure 3 presents the distribution of the corpus according to the number of words in the documents. The shortest document contains 190 words while the longest one contains 41,263 words. The average number of words per document is 3,687. Around 50% of the files present less than 1,024 words.

According to Komeno et al. (2015), a person who successfully finished the 9th grade in Brazil is able to silently read 196.14 words per minute on average. Considering a user with this level of literacy, it would take her 18.8 minutes to read an average-length privacy policy and up to 210 minutes to read the longest document in the corpus.

5.3 Readability

We use an adapted version of the Flesch Readability Score to evaluate the level of difficulty of the privacy policies (Martins et al., 1996). The Flesch version adapted to Portuguese returns an integer value between 0 and 100, which is then mapped to one of the four levels presented in Table 2. The third column of this table shows the minimum level of education one should have to understand the texts satisfactorily.

Figure 4 presents the distribution of policies according to its readability levels. No document scores more than 50 points. Thus, none of the policies is classified as easy or very easy. The average readability score is 19.11, that is, in general, files are very difficult to understand. The lowest and the highest readability values are -6.5 and 41.8, respectively.

¹https://github.com/valeriaquados/PPs_PT.

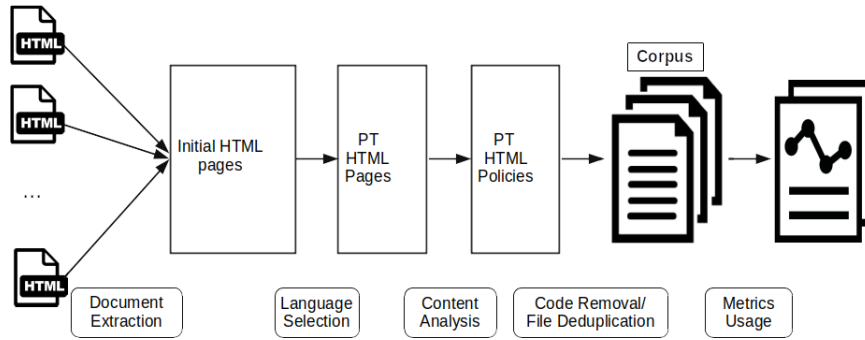


Figure 1. Steps performed during document collection, treatment and analysis.

Category	Description
First Party Collection/Use	How and why a service provider collects user information.
Third Party Sharing/Collection	How user information may be shared with or collected by third parties.
User Choice/Control	Choices and control options available to users.
User Access, Edit, & Deletion	If and how users may access, edit, or delete their information.
Data Retention	How long user information is stored.
Data Security	How user information is protected.
Policy Change	If and how users will be informed about changes to the privacy policy.
Do Not Track	If and how Do Not Track signals for online tracking and advertising are honored.
International & Specific Audiences	Practices that pertain only to a specific group of users (e.g., children, Europeans).
Other	Practices not covered by the other categories.

Table 1. Categories of data practices annotated in the corpus.

Flesch Value	Reading Difficulty	Level of Education
75-100	Very easy	1-4 ^o
50-75	Easy	5-9 ^o
25-50	Difficult	9-11 ^o
0-25	Very difficult	Graduated

Table 2. Flesch readability scores classified into four classes and the corresponding level of education (Martins et al., 1996).

That means all policies require a high degree of literacy from the reader to be understood (at least 9 years of study or higher level education).

The app *Conecta SUS*, provided by the Brazilian Ministry of Health, has the lowest readability score among all the privacy policies in the corpus. This document presents an average of 2.65 syllables per word and 30.8 words per sentence. This ratio of syllables per word is higher than the average typically found in texts written in Portuguese – 2.2 syllables per word in experiments conducted by Martins et al. (1996) –, and justifies the score extrapolation (-6.5) below the minimum expected value. In *Conecta SUS*, citizens can view their interactions with the Brazilian Universal Healthcare System (in Portuguese, Sistema Único de Saúde – SUS), such as exams, vaccines, and medication. The app is used by several people who depend exclusively on SUS for health care. Many of these people have low education and, consequently, would face great difficulties to understand the app’s privacy policy. The lack of knowledge about data treatment becomes even more critical when we consider that the information handled by *Conecta SUS* contains highly sensitive data.

An e-book app, called *Storytel*, presented the policy with the best readability (41.8). We hypothesized that certain categories of apps, such as those made for children, would present policies that are easier to understand. However, this

assumption is not confirmed. The policy of the game producer *Nintendo*, for example, obtained an index of -2.39 due to its average word length (2.58 syllables per word) substantially greater than the average of words traditionally used in Brazilian texts.

5.4 Privacy Policy Contents

APP-BR is composed of 6,650 segments. The vast majority of them (2,921 or 38.6%) contain none of the specific labeled categories, thus they are annotated as *Other*. Table 3 shows some descriptive statistics of the corpus. The high frequency

Category	Segments (abs.)	Segments coverage (%)	PPs coverage (%)
Other	2,921	38.6	100
First Party Collection/Use	1,718	22.7	100
Third Party Sharing/Collection	783	10.35	98.78
User Choice/Control	569	7.52	93.9
User Access, Edit, & Deletion	511	6.75	86.59
International & Specific Audiences	372	4.92	70.73
Data Security	307	4.06	91.46
Data Retention	193	2.55	73.17
Policy Change	176	2.33	95.12
Do Not Track	18	0.24	18.29

Table 3. By-category description for the data practices in the corpus.

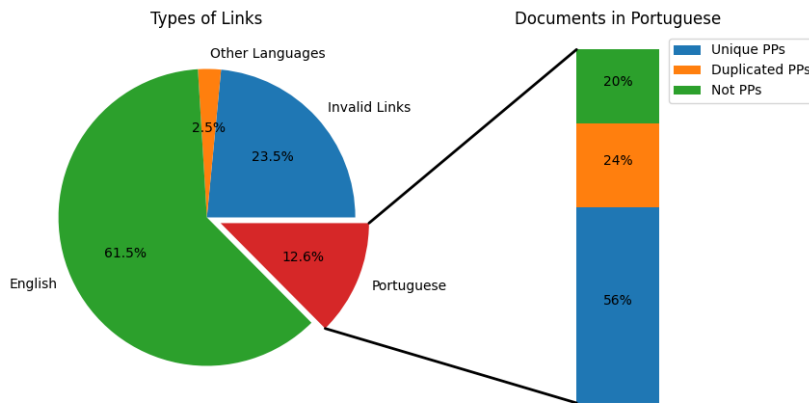


Figure 2. Fraction of links and files according to their types.

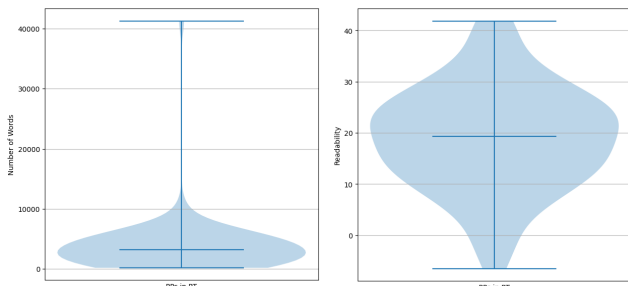


Figure 3. Distribution of policies according to number of words.

Figure 4. Distribution of policies according to readability level.

of the *Other* category raises two hypotheses:

- The amount of text could be summarized. Many paragraphs are just introductory and do not mention a specific practice.
- The list of data practices should be increased. New privacy concerns, such as the right to not have data used in automatic decision-making, are identified during the content analysis. Such segments should be labeled accordingly in future works.

The category *First Party Collection/Use* occurs in 1,718 segments in the corpus. This category is crucial, since it states what type of information the app provider collects and for what purposes. In the absence of data collection, none of the other data practices could be implemented. *Third Party Sharing/Collection* is present in 10.35% of the segments and in almost all the documents. Occasionally, this category and *First Party Collection/Use* exist in the same paragraph.

Rarely, *User Choice/Control* and *User Access, Edit, & Deletion* are also found in the same text chunk. These categories assure users the right to control and access their data. Despite their importance, 5 PPs do not address the first category, while 11 do not reference the second one. More than one quarter of the PPs (26.8%) do not inform for how long they store user data or the purpose of its retention. The ones that do report this feature only a comparably short text snippet on that matter (*Data Retention* paragraphs cover only 2.55% of the segments).

Data Security is present in more than 90% of the PPs. Segments of this category describe mechanisms for protection of data during user authentication, access, data transfers and

storage. *Policy Change* statements are identified in 176 segments, covering more than 95% of the PPs. *Do Not Track* was the least represented class. It is present in only 18 segments, which are distributed over 15 PPs. In fact, this type of privacy feature is rarely adopted in the industry.²

The results obtained with the annotation differ slightly from (Wilson et al., 2016). In OPP-115, the most popular categories are *First Party Collection/Use*, *Third Party Collection/Use*, *Other*, *User Choice/Control*, *Data Security*, *International & Specific Audiences*, *User Access, Edit, & Deletion*, *Policy Change*, *Data Retention* and *Do Not Track*. Understanding the causes of this discrepancy is out of the scope of this work. However, we envisage that Portuguese verbosity, differences in legislations and a stricter annotation process are likely explanations for this observation. Segment coverage is quite similar across the experiments.

Figure 5 presents the word cloud built upon the corpus contents. There are many references for words related to service, data, information, privacy and users. In his work, Pontes (2016) presents similar words as the most used ones.



Figure 5. Most used words in the corpus.

6 Privacy Policies and LGPD Compliance

The Brazilian General Data Protection Law regulates the collection and use of personal data since 2020 (L13709, 2018)

²<https://www.w3.org/2011/tracking-protection/>

and is composed of general rules regarding the rights of data subjects and the legal bases for the processing of personal data. Some of these rules are highly related to the labels used to annotate our corpus (see Table 1).

According to the LGPD, organizations can use personal data for licit purposes or legitimate interest. They may retain personal data only under envisaged legal basis and with reasonable security measures to protect users data. Considering these prerogatives, paragraphs informing about what type of data is collected, how and for what reasons are identified in APP-BR as *First Party Collection/Use*. Paragraphs regarding data preservation and security are tagged as *Data Retention* and *Data Security* practices, respectively.

The Brazilian LGPD states that individuals whose data is collected or processed have the right to:

1. Confirm that their personal data is being processed;
2. Access their personal data;
3. Correct incomplete, incorrect or out-of-date personal data;
4. Have anonymized, blocked, or deleted any unnecessary, excessive, or non-compliant personal data;
5. Request that a data controller move their personal data to another service or product provider (data portability);
6. Delete their personal data (with some exceptions where the data retention is necessary);
7. Be given information on public or private entities with whom, and how their personal data has been shared;
8. Be given information about their rights to not give consent to process their personal data, and consequences of refusal;
9. Revoke consent to process their personal data.

The first six items are clearly related to the *User Access, Edit & Deletion* data practice. Item 7 is covered by practice *Third Party Sharing/Collection*, while items 8 and 9 are covered by practice *User Choice/Control*.

Brazilian law foresees that changes in data processing policy must be publicly available. In case of a disagreement with the new policy, users can revoke their consent about personal data usage. The explicit discussion of such aspects in the PPs were labeled as *Policy Changes* practices.

The protection law imposes no hard *Do not Track* agreement between users and service providers. However, it has special provisions for children and their data, in a basis similar to the proposed *International & Specific Audience* practice. Last but not least, the LGPD includes situations that do not fit in any of the data practices considered in this work. Some relevant examples are:

- The right to request a review of decisions taken solely on the basis of automated processing of personal data;
- The need for specific treatments when dealing with sensitive data or international data transfers;
- The mandatory exposition of the responsible for the data treatment.

Considering all this and also considering that *First Party Collection/Use* was the only practice present in 100% of the policy files, one can confidently affirm that Brazilian privacy policies are not LGPD compliant, mainly in terms of data retention and children's information.

7 Threats to Validity

As presented in Section 4.2, we used an annotation process already established and validated for semantic analysis. Other issues were considered to ensure the control of our research, such as: i) subsets of documents were randomly assigned to collaborators, so to minimize differences among them; and ii) the annotators worked independently. Considering the criteria and the methodology adopted, we infer that our experimentation poses a satisfactory degree of internal validity. This is evidenced by the similarity of our results with the results obtained in (Wilson et al., 2016). Even so, we cite some weaknesses of our study.

Text processing depends highly on the accuracy of the employed computational tools. The boilerplate removal methods evaluated in this work showed large variations in quality, and none of them could deliver perfect results. Paragraph segmentation was also done in the best effort manner. More reliable data could be manually extracted and chunked from files. However, for the sake of scalability and reproducibility, we decided to automate this step of the experiments.

Text annotation involves a lot of subjectivity. Some paragraphs are not clear about the data practices they refer to. Frequently, *First Party Collection/Use* is implicit in *Third Party Collection/Use* practices (and vice-versa). The same occurs with the pair *User Choice/Control* and *User Access, Edit & Deletion*. Subjectivity is also present in the legislation interpretation.

An alternative to reduce threats in the annotation process is quite expensive. It would employ specialists in law to propose a custom set of data practices in conformance to the LGPD, and then perform a long annotation training using the data labels scheme. Afterwards, each paragraph should be annotated by a group of experts before the results are consolidated.

In terms of external validity, since our corpus is composed of a wide variety of categories of applications – ranging from games to educational apps –, we presume that our analysis is suitable to be applied to new datasets.

8 Sociological Issues

From our analysis, we can affirm that mobile apps in Brazil often neglect their users' right to access privacy policies. PPs are frequently very difficult to understand, incomplete, or even non-existent. According to Continuous National Household Sample Survey (in Portuguese, Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua), in 2019, 11 million Brazilians over 15 years old, corresponding to 6.6% of the population, were illiterate (IBGE, 2019). The same survey evidenced that only 48.8% of people aged 25 or more have completed higher education. These rates increase as the age group advances. The illiteracy rate among elderly people is 18%.

In 2019, the 5th edition of the Reading Portraits survey in Brazil (in Portuguese, Retratos da Leitura) revealed that 48% of the Brazilian population does not have the habit of reading (IBOPE Inteligência, 2019). This fact, combined with the population's lack of digital education, exacerbates the

problem of awareness of privacy in the digital world (Soares et al., 2020). Another aspect that should be considered is the number of people with conditions that negatively impacts their skills to read and understand texts. People with some degree of visual impairment correspond to 3.4% of the Brazilian population, and people with some degree of mental/intellectual disability are 1.4% of the population (IBGE, 2010).

Another relevant piece of data, provided by the British Council, informs that, in 2013, only 5.1% of the Brazilian population aged 16 or over had some knowledge of the English language (Council, 2014). Considering that many privacy policies for apps available in Brazil are provided only in this language, this is an important issue.

Regarding the presented statistics, we conclude that texts with low level of readability will not be fully understood by a considerable percentage of the Brazilian population. Considering that privacy policies are of public interest, it is important that they present good readability. Thus, it is necessary to invest more in the creation of easily readable contracts as well as in digital education for the population.

Another violation of citizen's rights is the lack of information regarding some topics in the LGPD. Our analysis opened the discussion about this theme and motivates the development of mechanisms to improve law conformance in PPs. Besides the population, these mechanisms could benefit also digital service providers and regulatory agencies.

9 Conclusion

In this work, we described the creation of a corpus composed of privacy policies written in Portuguese. We also presented a vast analyses considering quantitative and qualitative aspects. To the best of our knowledge, this is the first work of this nature that considers Brazilian privacy policies.

In the creation process, we used computational tools to collect and process data. We manually analyzed the collected files to consider only Portuguese privacy policies and to characterize/classify their contents. Out of 1,163 collected documents, only 82 composed the corpus described in this work, given that all the other documents were not privacy policies or not written in Portuguese.

Initially, the corpus was characterized using 2 metrics: the number of words and the level of readability. In general, privacy policies were very extensive and complex, even for people with a high level of education. This negative aspect is in line with existing studies for other languages.

Regarding the content of the policies, our analysis uncovered the main topics treated in these documents. We found that 40% of the segments make generic statements without describing any specific data practice considered in the label set. As expected, *Data Collection/Use* was the most common practice mentioned in the corpus, being present in all documents. Other practices, however, were not covered at all in some privacy policies. This lack of information makes some app providers in non-conformance with the Brazilian General Data Protection Law. The worst covered categories were *Data Retention* and *International & Specific Audiences*.

We claim that many of the privacy policies from mobile

apps fail to inform the public about data practices. In this sense, our conclusions contribute to the advancement in the construction of clear, objective and law compliant digital contracts.

As future works, we plan to explore the knowledge provided by our annotated corpus. Among the possibilities, we highlight the implementation of machine learning models to extract key data usage practices described in the documents. This could render the creation of tools for summarization possible, as well as automate new privacy policies annotations. Another possibility is to exploit similarities of policies contained in APP-BR and OPP115, to merge them into a bigger data set using automatic translation. Finally, we also intend to use all the learnings gained from building the corpus to devise a new data set of annotated policies. The new set would comprise all the data handling practices foreseen by LGPD.

Acknowledgements

The present work was supported by the Federal University of Mato Grosso do Sul (UFMS), Brazil, and Leuphana Universität of Lüneburg Germany.

References

- Abreu, J. (2018). As políticas de privacidade de apps do governo. <https://internetlab.org.br/pt/noticias/especial-as-politicas-de-privacidade-de-apps-do-governo/>. [Online: last access in 17-3-2022].
- Barbosa, P. H. M. (2017). Análise das permissões e violações de privacidade em aplicações para android. <https://www.cin.ufpe.br/~tg/2017-2/phmb2-tg.pdf>. [Online: last access in em 28-1-2022].
- Barboza, E. M. F. and Nunes, E. M. d. A. (2008). A inteligibilidade dos websites governamentais brasileiros e o acesso para usuários com baixo nível de escolaridade. *Inclusão Social*, 2(2).
- Berger-Walliser, G., Barton, T. D., and Haapio, H. (2017). From visualization to legal design: A collaborative and creative process. *American Business Law Journal*, 54(2):347–392.
- Council, B. (2014). Demandas de aprendizagem de inglês no brasil. https://www.britishcouncil.org.br/sites/default/files/demandas_de_aprendizagem_pesquisacompleta.pdf. [Online: last access in 28-1-2022].
- d'Aquin, M., Kirrane, S., Villata, S., Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J., Sadeh, N., d'Aquin, M., Kirrane, S., and Villata, S. (2018). Privonto: A semantic framework for the analysis of privacy policies. *Semant. Web*, 9(2):185–203.
- Fabian, B., Ermakova, T., and Lentz, T. (2017). Large-scale readability analysis of privacy policies. In *WI '17: Proceedings of the International Conference on Web Intelligence*, New York, NY, USA. Association for Computing Machinery.

- Flesch, R. (1979). *How to write plain English: a book for lawyers and consumers*. Harper & Row.
- IBGE (2010). Censo demográfico 2010. https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf. [Online: last access in 28-1-2022].
- IBGE (2019). PNAD Educação 2019. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf. [Online: last access in 28-1-2022].
- IBOPE Inteligência (2019). Retratos da leitura no brasil. https://prolivro.org.br/wp-content/uploads/2020/09/5a_edicao_Retratos_da_Leitura_no_Brasil_IPL-compactado.pdf. [Online: last access in 28-1-2022].
- Jardim, G., Rabello, M., Lima, A., Brefeld, U., and Reis, V. (2022). Uma caracterização das políticas de privacidade utilizadas em aplicativos no brasil. In *Anais do III Workshop sobre as Implicações da Computação na Sociedade*, pages 13–25. SBC.
- Komeno, E. M., de Ávila, C. R. B., de Pádua Cintra, I., and Schoen, T. H. (2015). Velocidade de leitura e desempenho escolar na última série do ensino fundamental. *Estudos de Psicologia*, 32(3):437–447.
- L13709 (2018). Lei Nº 13.709 de 14 de agosto de 2018: Lei geral de proteção de dados pessoais (LGPD). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm.
- Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. d. G. V., and Oliveira Junior, O. N. d. (1996). Readability formulas applied to textbooks in brazilian portuguese. Technical report, ICMCS-USP.
- Obar, J. A. and Oeldorf-Hirsch, A. (2018). The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services. In Information, C. . S., editor, *TPRC 44: The 44th Research Conference on Communication, Information and Internet Policy*, pages 1–20.
- Pollach, I. (2007). What’s wrong with online privacy policies? *Commun. ACM*, 50(9):103–108.
- Pontes, D. R. G. d. (2016). Geração de rótulo de privacidade por palavras-chaves e casamento de padrões. <https://repositorio.ufscar.br/bitstream/handle/ufscar/8730/DissDRGP.pdf>. [Online: last access in 28-1-2022].
- Siebra, S. d. A. and Xavier, G. A. C. (2020). Políticas de privacidade da informação: caracterização e avaliação. *BIBLOS*, 34(2).
- Soares, H. J., Araújo, N. V. d. S., and de Souza, P. (2020). Privacidade e segurança digital: um estudo sobre a percepção e o comportamento dos usuários sob a perspectiva do paradoxo da privacidade. In *Anais do I Workshop sobre as Implicações da Computação na Sociedade*, pages 97–106. SBC.
- Viana, G. T., Maciel, C., de Arruda, N. A., and de Souza, P. C. (2017). Análise dos termos de uso e políticas de privacidade de redes sociais quanto ao tratamento da morte dos usuários. In *Anais do VIII Workshop sobre Aspectos da Interação Humano-Computador para a Web Social*, pages 82–93, Porto Alegre, RS, Brasil. SBC.
- Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Giovanni Leon, P., Schaarup Andersen, M., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., and Sadeh, N. (2016). The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.
- Wilson, S., Schaub, F., Liu, F., Sathyendra, K. M., Smullen, D., Zimmeck, S., Ramanath, R., Story, P., Liu, F., Sadeh, N., and Smith, N. A. (2018). Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1).
- Yamauchi, E. A., Souza, P. C. d., and Junior, D. (2016). Questões proeminentes para o estabelecimento da privacidade em políticas de privacidade de app móveis. In *XV Brazilian Symposium on Human Factors in Computing Systems (IHC 2016)*.