Uma Caracterização das Políticas de Privacidade Utilizadas em Aplicativos no Brasil

Guilherme P. S. Jardim[1], Maria E. R. Rabello[2], Anderson C. Lima[2], Ulf Brefeld[3], Valéria Q. Reis[2]

[1]Universidade Estadual de Campinas

[2]Universidade Federal do Mato Grosso do Sul

[3]Universität Leuphana

guilhermepsjardim@gmail.com, {elisa.rabello,anderson.lima}@ufms.br

brefeld@leuphana.de, valeria.reis@ufms.br

Resumo. *Políticas de privacidade são documentos nos quais empresas de tecnologia especificam como os dados de seus usuários são tratados. No Brasil, a concordância entre essas duas partes está prevista em lei. Assim, é fundamental para os usuários que os documentos disponibilizados sejam facilmente acessados e compreendidos. Dessa maneira, o objetivo deste trabalho foi caracterizar o acesso à informação em documentos de privacidade de mais de 1.000 aplicativos da Google Play Store. Os resultados mostraram que muitos aplicativos apresentavam documentos inválidos e somente 10% disponibilizavam políticas de privacidade escritas em português. Ademais, nenhum desses documentos dispunha de um grau de inteligibilidade adequado a boa parte da população brasileira.*

Abstract. *Privacy policies are documents in which technology companies specify how personal data from their users are processed. In Brazil, the agreement between companies and users is regulated by law. So, it is important to users to be able to easily access and understand such documents. In this work, we study privacy policies from more than 1,000 apps from Google Play Store. We show that many of these apps provide invalid documents and only approximately 10% of them contain privacy policies in Portuguese. Furthermore, none of the considered documents have an adequate degree of readability and will be incomprehensible for large parts of the Brazilian population.*

1. Introdução

Desde 2020 está em vigor no Brasil a Lei Geral de Proteção de Dados (LGPD), a qual dispõe sobre a proteção de dados pessoais, visando preservar os direitos fundamentais de liberdade e de privacidade dos brasileiros [Lei Nº 13.709 2021]. Segundo a LGPD, os serviços digitais, tais como aplicativos e sites da Internet, devem obter explicitamente o consentimento de seus usuários para o tratamento de dados. Devido a essa obrigatoriedade de concordância entre as partes, políticas de privacidade se popularizaram como instrumentos de informação sobre a manipulação de dados pessoais pelas empresas de tecnologia [Siebra and Xavier 2020].

Contudo, o acesso às políticas de privacidade ainda é negligenciado quando se considera a existência de instituições que não tornaram públicas suas práticas de uso de dados. Um estudo, realizado em 2018, analisou 13 aplicativos do governo federal e revelou que pelo menos seis deles não possuíam política de privacidade acessível publicamente [Abreu 2018].

Políticas de privacidade incompletas, ou seja, que não listam todos os dados tratados pelas empresas ou o motivo do tratamento, também são obstáculos para a instrução dos usuários. Em 2017, uma análise de aplicações da Google Play Store detectou que, após atualizações de código, variações na quantidade de permissões das aplicações ocorriam sem que essas alterações se refletissem nas políticas de privacidade. O mesmo estudo também analisou códigos de aplicações do GitHub e encontrou incompatibilidades entre as práticas de uso de dados implementadas e aquelas descritas nas respectivas políticas de privacidade [Barbosa 2017].

Uma última complicação no acesso à informação surge quando a política de privacidade é de difícil entendimento. Problemas assim ocorrem devido ao tamanho excessivo dos textos, ao uso de linguagem rebuscada ou à ambiguidade intencional. Muitos trabalhos abordam a dificuldade de compreensão de políticas no contexto da língua inglesa. Um exemplo é dado no trabalho de [Pollach 2007], em que após análise de 50 políticas de privacidade de sites norte americanos, concluiu-se que tais documentos careciam de conteúdo centrado nos interesses dos usuários, assim como de formatos de apresentação mais amigáveis. Ainda no cenário americano, em 2017, métricas de inteligibilidade foram aplicadas em quase 50.000 políticas de privacidade da Web [Fabian et al. 2017]. Os resultados mostraram que, em média, políticas de privacidade exigem um alto nível de entendimento, sendo necessários entre 9 e 15 anos de estudo para compreendê-las.

Os problemas encontrados nas políticas de privacidade desestimulam a leitura e dificultam a compreensão das práticas de dados. Consequentemente, muitas pessoas não leem as políticas de privacidade antes de acessarem os serviços digitais de que necessitam [Obar and Oeldorf-Hirsch 2018]. Ainda não são observados no contexto brasileiro ou da língua portuguesa trabalhos consolidados que realizaram estudos sobre as características das políticas de privacidade. Esta lacuna será abordada neste trabalho a partir da criação e da caracterização de um *corpus* de políticas de privacidade escritas em língua portuguesa e disponibilizadas por aplicativos no Brasil.

A criação do *corpus* proposto é importante porque torna os padrões de acesso às políticas mais claros, destacando qual parcela dos serviços apresenta práticas de privacidade e identificando os obstáculos que o usuário enfrenta para obter essas informações. A partir do *corpus*, é possível otimizar métricas diretamente relacionadas à compreensão dos documentos. Por fim, um *corpus* é fundamental para a implementação de serviços que extraiam automaticamente informações das políticas e as apresentem em formatos mais amigáveis para o usuário, tais como resumos e textos com perguntas e respostas.

Para compor o *corpus* foram coletadas 1.163 políticas de privacidade dos aplicativos mais baixados da *Google Play Store* no Brasil. Após o tratamento dos documentos coletados, a caracterização do *corpus* foi feita de forma manual para a classificação de textos. Para a obtenção de estatísticas de conteúdo e métricas de inteligibilidade utilizamos bibliotecas específicas de linguagem natural.

O restante deste artigo está estruturado como segue: a Seção 2 apresenta conceitos teóricos utilizados ao longo do texto e os trabalhos relacionados; a Seção 3 descreve os passos necessários para a construção do *corpus* e para a sua análise; a Seção 4 expõe os resultados obtidos desta análise; a Seção 6 apresenta discussões sobre os impactos indiretos desses resultados na sociedade; e, por fim, a Seção 7 apresenta as conclusões do trabalho e elenca oportunidades de pesquisas futuras.

2. Fundamentação Teórica

Para embasar a pesquisa, foram utilizados artigos que pudessem ajudar na compreensão do panorama atual sobre políticas de privacidade.

2.1. As Políticas de Privacidade e os Termos de Uso

Os serviços digitais devem dispor de Políticas de Privacidade sempre que houver interação entre usuários e sistemas. Políticas de privacidade regulam o uso de dados pessoais pelas empresas de tecnologia [Siebra and Xavier 2020]. Através delas, os usuários dos serviços digitais podem avaliar e julgar como seus dados pessoais são tratados e, a partir de tal julgamento, decidir se aceitam as regras impostas no contrato. Dessa maneira, políticas de privacidade devem determinar quais dados do usuário são coletados, de que forma, por qual motivo, como são armazenados e por quanto tempo. Os documentos ainda devem informar se os dados são compartilhados com terceiros e se o usuário e empresa/governo podem alterar suas opções de concordância a qualquer momento.

Frequentemente, Termos de Uso são confundidos com políticas de privacidade, mas diferem delas ao estabelecerem normas de utilização do serviço pelo usuário e delinearem as responsabilidades das empresas [Yamauchi et al. 2016]. Na caracterização realizada neste trabalho, são consideradas somente políticas de privacidade.

2.2. O Design Legal

O *Design Legal* é a área que se dedica ao estudo de como documentos jurídicos deveriam ser de modo a se tornarem mais claros e diretos para as pessoas [Berger-Walliser et al. 2017]. Exemplos de boas práticas do *Design Legal* são textos resumidos, uso de termos de fácil entendimento, criação de perguntas e respostas, e exposição de cláusulas contratuais através do uso de imagens.

Situações em que os *links* de políticas de privacidade dos aplicativos levam a documentos que não são as práticas de privacidade, tais como páginas de propaganda, consistem em exemplos de práticas ruins de *design legal*, mesmo quando tais páginas apresentam *links* para a política de interesse. O uso de *captchas* e *cookies* também é configurado como prática indevida. *Captchas* dificultam o acesso à informação e *cookies* envolvem a manipulação de informações do usuário e, por esse motivo, devem eles próprios serem dispostos no próprio corpo da política de privacidade.

2.3. Inteligibilidade

Inteligibilidade consiste em uma medida de quão confortável ou facilmente um texto pode ser lido. Com frequência, esse termo é confundido com legibilidade, o qual refere-se à apresentação gráfica do texto, tais como a diagramação do conteúdo e o tamanho/desenho das fontes. Portanto, a legibilidade é parte importante da inteligibilidade, mas o inverso

não se aplica. Um documento pode ser legível, porém não inteligível. A inteligibilidade depende ainda da habilidade de leitura do usuário. Um documento jurídico, por exemplo, é mais facilmente compreendido por um especialista do que por pessoas com baixo nível de instrução [Barboza and Nunes 2008].

O índice Flesch é uma métrica de inteligibilidade amplamente utilizada na comunidade científica, sendo a mesma já adaptada e validada para a língua portuguesa [Flesch 1979, Martins et al. 1996]. A adaptação do índice ao português considera que a pontuação obtida por textos neste idioma são usualmente maiores do que as pontuações dos textos em língua inglesa. Sua fórmula é dada pela equação a seguir, onde MPF é o tamanho médio das frases (número de palavras dividido pelo número de frases) e MSP é o tamanho médio das palavras (número de sílabas dividido pelo número de palavras).

$$INTELIG = 248,835 - (1,015 * MPF) - (84,6 * MSP)$$

O valor de inteligibilidade tende a variar entre 0 e 100, mas eventualmente, pode extrapolar tais valores quando os padrões do texto diferem significativamente da média esperada. Quanto maior o valor de inteligibilidade, mais fácil de ler é o texto considerado. Os valores encontrados são interpretados em uma escala de 4 níveis: muito difícil, difícil, fácil e muito fácil.

2.4. Trabalhos Correlatos

Internacionalmente, destaca-se o estudo conduzido em [Obar and Oeldorf-Hirsch 2018], que apresenta através de um experimento com 543 participantes, as razões pelas quais indivíduos ignoram a leitura das políticas de privacidade. Entre elas estão o tamanho excessivo dos textos, o sentimento de que seus dados não precisam ser privados, a necessidade de uso dos serviços digitais a qualquer preço e a dificuldade de entendimento dos documentos. Este trabalho corrobora a primeira hipótese levantada pelos autores.

A causa da dificuldade de entendimento de políticas de privacidade é abordada no artigo [Fabian et al. 2017], que detalha a implementação de uma ferramenta de extração e análise de políticas de privacidade. No trabalho, foram obtidas as políticas de privacidade dos 202.144 sites mais populares segundo o serviço Alexa[1]. Destes, apenas 163.232 páginas possuíam conteúdo em inglês, e, de acordo com um algoritmo de classificação desenvolvido, apenas 1 em cada 3 páginas consistia de fato em uma política de privacidade. As análises das 49.036 políticas encontradas revelaram que os documentos possuíam tamanho médio de 1.700 palavras e índices de inteligibilidade de difícil compreensão, pois, em geral, requeriam um nível de instrução superior ao do ensino médio americano. Outra contribuição importante do trabalho foi apresentar a forte correlação do índice Flesch com outros índices amplamente utilizados na literatura. A análise de Fabian et al. [Fabian et al. 2017] apresentou metodologia e resultados similares aos realizados neste trabalho, além de validar o índice que inteligibilidade utilizado.

Por fim, ressalta-se, em 2016, a apresentação do *corpus* 0PP-115. Trata-se de um conjunto valioso de dados oriundos de 115 políticas de privacidade de *websites* norte-americanos em língua inglesa. Os autores afirmam que se tratou do primeiro trabalho

[1]https://www.alexa.com/topsites/

em grande escala para a anotação de políticas de privacidade em um nível refinado de detalhes. O trabalho destacou que o OPP-115 foi um passo importante na pesquisa de métodos automatizados de anotação de políticas de privacidade, mas ressaltou que ainda persistiam lacunas na forma como as políticas de privacidade eram apresentadas e compreendidas pelos usuários [Wilson et al. 2016]. Tais lacunas também foram identificadas nas análises deste trabalho e igualmente vislumbra-se a criação de um *corpus* anotado de políticas em português.

No cenário nacional, não é de nosso conhecimento a existência de trabalhos que analisem a inteligibilidade de políticas de privacidade. No entanto, há de se destacar o trabalho de [Pontes 2016] que, com o uso de técnicas de mineração de dados, sintetizou um conjunto de 50 políticas de privacidade categorizadas em formato tabular, denominadas como rótulos. O trabalho descreve que rótulos podem ser mais facilmente compreendidos pelos usuários, economizando tempo de leitura e removendo terminologias jurídicas complicadas. O método consistiu na seleção do *corpus* oriundo das políticas de privacidade dos 60 *websites* mais acessados no Brasil. Em seguida, especialistas do domínio utilizaram um protótipo semi-automático para a geração e a análise de rótulos.

Outras iniciativas se dedicaram à análise de permissões e violações de privacidade em aplicativos móveis. Por exemplo, um trabalho de 2017 verificou 41 aplicações Android e 10 aplicações de código aberto. Os resultados apontaram variações significativas na quantidade de permissões a cada atualização de versão, além de muitas violações de privacidade nas aplicações com código disponibilizado [Barbosa 2017]. Em outra frente, um trabalho de 2020 propôs um conjunto de dez diretrizes para orientar projetistas e desenvolvedores no projeto de interfaces de privacidade mais compreensíveis [Yamauchi et al. 2016]. Ainda em 2020, outro trabalho apresentou uma lista de 17 critérios para avaliar a qualidade e completude de políticas de privacidade [Siebra and Xavier 2020].

A falta de consciência dos usuários em relação à segurança e privacidade digital foi tema abordado no trabalho de Souza et al. [Soares et al. 2020]. Através de duas pesquisas exploratórias, os autores concluíram ser necessário mais investimentos na educação digital, mesmo para públicos com conhecimento prévio em computação.

3. Procedimento Metodológico

O procedimento metodológico contemplou etapas de coleta, pré-processamento e análise das políticas. A Figura 1 ilustra fases desse procedimento. Devido ao grande volume de dados a serem obtidos e visando a facilidade na replicação dos experimentos, essas fases foram, em grande parte, automatizadas.

De modo a estabelecer um conjunto de políticas que satisfizesse algum indicador de relevância social, optou-se pela coleta de políticas dos aplicativos mais populares da Google Play Store [2] no mês de fevereiro de 2021. A Google Play Store é a maior loja de aplicativos no segmento de dispositivos móveis no Brasil.

A partir das páginas dos aplicativos mais populares da loja virtual, *web crawlers* salvaram os arquivos apontados pelos *links* das políticas de privacidade. A coleta resultou em um conjunto de 1.163 páginas HTML.

[2]https://play.google.com

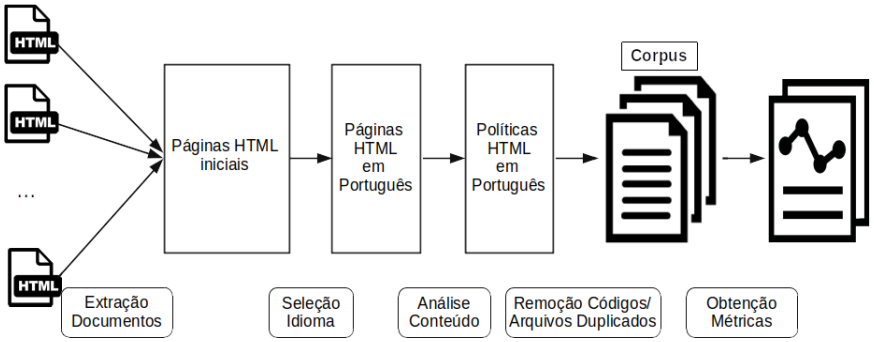

Figura 1. Subfases de coleta, tratamento e análise de documentos.

Iniciou-se então a detecção automática dos idiomas dos arquivos com o intuito de excluir políticas que não estivessem escritas em português. Alguns erros de classificação ocorreram e tiveram que ser corrigidos por uma revisão manual, em que se observaram muitos arquivos inválidos ou que não possuíam relação com práticas de uso de dados. Tais arquivos foram removidos do conjunto final, assim como arquivos que descreviam de maneira incompleta o tratamento de dados. A revisão serviu ainda para identificar documentos que continham características positivas – escolha de idioma, formato de perguntas e respostas – e negativas – uso de *captchas*, *cookies* – em relação ao acesso à informação.

Com a base de dados sem arquivos duplicados, removeu-se, por meio da ferramenta Trafilatura[3], os códigos HTML, CSS e JavaScript dos arquivos de modo a garantir que apenas o conteúdo das políticas fosse utilizado na caracterização [4].

Os arquivos texto foram submetidos à ferramenta PyLinguistics[5] para a obtenção do número de palavras e do índice Flesch de cada documento. Por fim, sobre a luz dos números obtidos, foi realizada uma análise descritiva do *corpus*.

4. Análise dos Resultados

Dos 1.163 aplicativos considerados neste trabalho, apenas 926 possuíam *links* válidos. Dentre os documentos coletados a partir destes links, somente 146, ou seja, 12,6% estavam escritos em português. Grande parte das supostas políticas de privacidade estavam escritas em inglês, mais especificamente 715 delas ou 61,5%.

No lado esquerdo da Figura 2, ilustram-se as porcentagens de documentos em cada grupo descrito (*links* inválidos, documentos em português, inglês e outros idiomas). Do lado direito, há uma representação dos 146 documentos escritos em português. Entre esses documentos:

- 29 foram desconsiderados (20%), visto que se tratavam de termos de uso (9), práticas incompletas de tratamento de dados (2), informações diversas (13), ou outros problemas;
- 26 exigiam concordância com o uso de *cookies* para obter acesso ao teor do arquivo;

[3]https://trafilatura.readthedocs.io/en/latest/evaluation.html
[4]Corpus disponível em https://github.com/valeriaquadros/PPs_PT.git.
[5]https://github.com/vwoloszyn/pylinguistics

- 32 apresentavam informações no formato perguntas e respostas.

Após a exclusão do primeiro grupo listado, a base de dados passou a possuir 117 arquivos com políticas de privacidade (56% do conjunto de documentos em português). Esse número tornou-se ainda menor após a remoção de documentos replicados. O Facebook, por exemplo, publicou a mesma política para os aplicativos Whatsapp, Facebook e Instagram. Removendo os arquivos idênticos, obteve-se o *corpus* final com uma amostra de 82 políticas de privacidade em português.

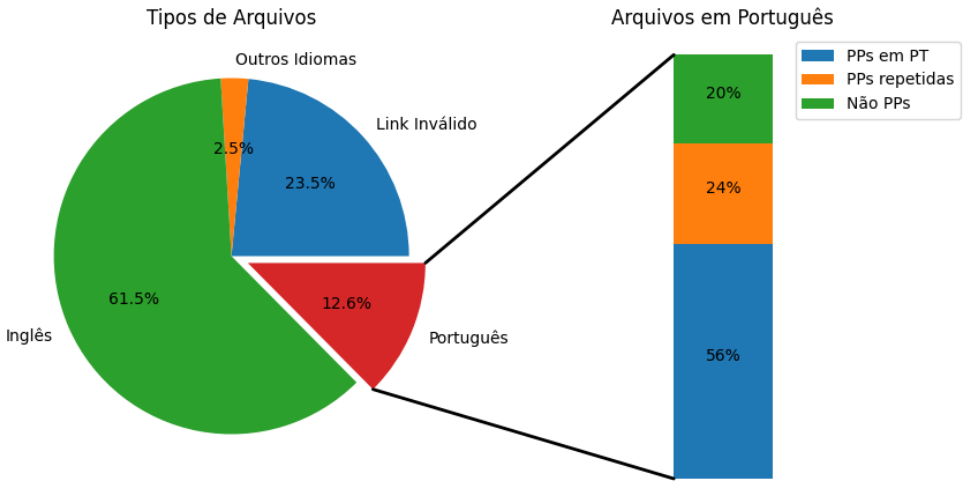

Figura 2. Divisão dos arquivos por tipo.

Conclui-se que, entre um conjunto de 1.163 documentos, apenas 117, ou seja, 10% correspondiam ao perfil buscado para compor o *corpus*, o que demonstra que usuários de meios digitais no Brasil têm dificuldade de acesso às práticas de uso de dados adotadas por empresas.

A seguir, apresenta-se a análise do *corpus* final em relação ao tamanho dos documentos e à inteligibilidade dos textos.

4.1. Tamanho dos Documentos

Na Figura 3, é apresentada a distribuição de arquivos conforme o número de palavras que eles contêm. O menor arquivo, correspondente ao aplicativo *Bíblia Narrada Cid Moreira*, possuía 190 palavras e o maior, correspondente à política padrão da Microsoft, possuía 41.263. A média de palavras observada foi de 3.687 por documento. Cerca de 50% dos documentos apresentaram um número de palavras menor que 1.024.

De acordo com [Komeno et al. 2015], a velocidade de leitura silenciosa de uma pessoa que estudou até o nono ano no Brasil é de 196,14 palavras por minuto. Considerando essa taxa, um usuário levaria 18,8 minutos para ler uma política de privacidade de tamanho médio e até 210 minutos para ler a maior política de privacidade encontrada.

Ainda há de se discutir o nível de dificuldade dos textos considerados. Nos experimentos de Komeno, foram utilizados textos com baixa dificuldade de compreensão. Neste estudo, como será apresentado a seguir, as políticas de privacidade analisadas apresentaram uma dificuldade de compreensão moderada.

4.2. Nível de Inteligibilidade

O teste de inteligibilidade utilizado é uma adaptação do índice Flesch para o português [Martins et al. 1996]. O valor de retorno desse teste consiste em um número mapeado para um dos quatro níveis de dificuldade apresentados na Tabela 1. Na terceira coluna dessa tabela há ainda a escolaridade esperada para que a compreensão do texto analisado seja satisfatória. Um texto com índice entre 75 e 100, por exemplo, é facilmente compreendido por uma pessoa, que não domina a leitura de forma satisfatória. Por outro lado, um texto cujo índice é inferior a 25 é mais facilmente compreendido por pessoas com nível superior ou que ao menos tenham completado o ensino médio.

Valor	Dificuldade de Leitura	Escolaridade
75-100	Muito fácil	1-4º ano
50-75	Fácil	5-9º ano
25-50	Difícil	9-11º ano
0-25	Muito difícil	Nível Superior

Tabela 1. Interpretação do teste de inteligibilidade Flesch.

Na Figura 4, apresenta-se a distribuição das políticas de acordo com seus níveis de inteligibilidade. Nota-se que todos os documentos obtiveram índice Flesch inferior a 50. Dessa maneira, nenhuma política foi classificada como sendo de fácil leitura.

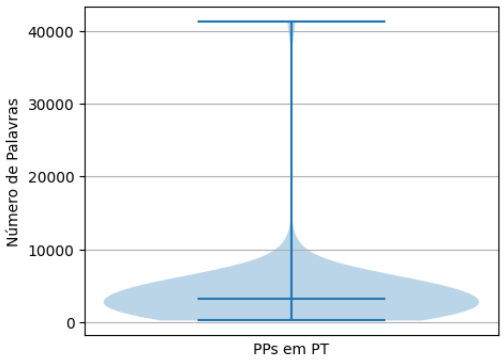

Figura 3. Distribuição das políticas por número de palavras.

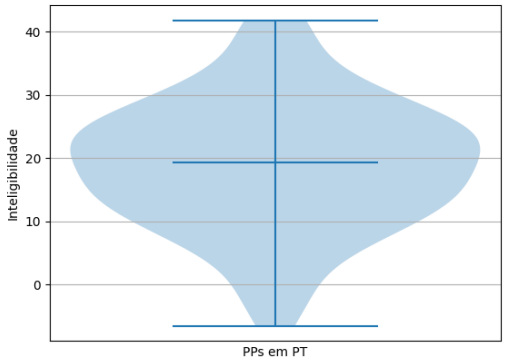

Figura 4. Distribuição das políticas por nível de inteligibilidade.

A média do nível de inteligibilidade encontrada foi de 19,11, um valor atribuído a textos muito difíceis de se ler. O menor valor de inteligibilidade encontrado foi -6,5 e o maior, 41,8, ou seja, todas as políticas exigiam um alto grau de compreensão, sendo indicadas para leitores com ao menos 9 anos de estudo ou nível superior.

A política do aplicativo *Conecta SUS*, disponibilizado pelo Portal do Ministério da Saúde, apresentou o menor nível de inteligibilidade (-6,5) do *corpus*. Essa política apresentava médias de 2,65 sílabas por palavra (ponto inferior mais à direita na primeira imagem da Figura 5) e 30,8 palavras por frase (ponto mais abaixo na segunda imagem da Figura 5). Esta relação de sílabas por palavra é maior que as frequentemente encontradas em textos da língua portuguesa e justifica o motivo do índice Flesch ter extrapolado o valor mínimo esperado.

No *Conecta SUS*, cidadãos podem visualizar as interações realizadas nos pontos de atenção à saúde e acompanhar o seu histórico no Sistema Único de Saúde (SUS), como exames, vacinas, dispensação de medicamentos e localização de estabelecimentos de atendimento à população. O aplicativo é utilizado por diversas pessoas que dependem exclusivamente do SUS para cuidados com a saúde, muitas das quais têm baixa escolaridade e, consequentemente, teriam dificuldade na compreensão da política de privacidade do aplicativo. A gravidade da situação torna-se ainda maior quando consideramos que as informações manipuladas pelo *Conecta SUS* contém dados altamente sensíveis.

Ironicamente, um aplicativo de livros eletrônicos foi o que apresentou a política com o melhor nível de inteligibilidade (41,8). A política do *Storytel* utiliza palavras com uma média de 2,18 sílabas e frases com uma média de 22,4 palavras.

Era esperado que determinadas categorias de aplicativos, tais como aquelas destinadas ao público infantil, apresentassem políticas de mais fácil compreensão, mas não foi esse o cenário observado. A política da produtora de jogos *Nintendo*, por exemplo, obteve índice -2,39, pois o tamanho médio de suas frases era de 2,58 palavras, substancialmente maior que a média das palavras tradicionalmente utilizadas em textos do Brasil.

A Figura 5 apresenta todos os valores de inteligibilidade segundo as médias de sílabas por palavra (esquerda) e palavras por frase (direita).

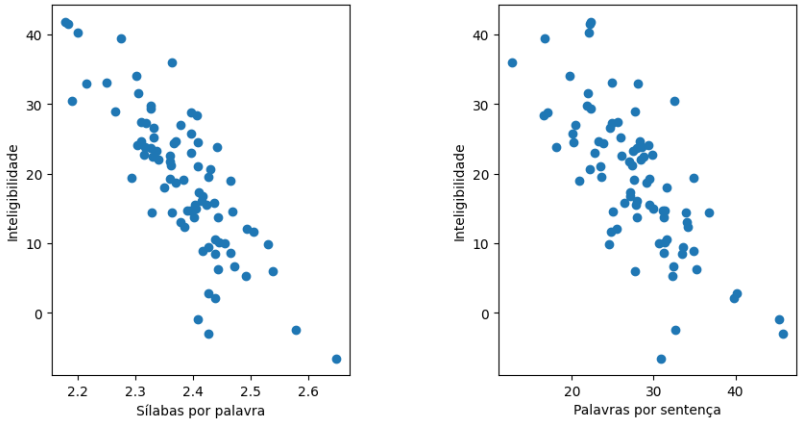

Figura 5. Nível de inteligibilidade segundo as médias de sílabas por palavra (esquerda) e palavras por frase (direita).

Na fórmula de inteligibilidade utilizada neste trabalho (Seção 2.3), não há relação direta entre o tamanho do arquivo e a sua inteligibilidade. Por exemplo, o menor arquivo de nossa base de dados, com 190 palavras, possuía 12,7 palavras por frase, 2,36 sílabas por palavra e obteve índice Flesch igual a 36,1.

5. Ameaças à Validade

Não há evidência empírica que suporte a imutabilidade das políticas. É esperado que o período no qual a coleta dos dados é realizada implique grande influência nos resultados, uma vez que as políticas tendem a sofrer alterações. Seria interessante conduzir medição estatística da sazonalidade de mudanças para melhor generalização de inferências.

O critério de seleção das políticas também é primordial para a obtenção de resultados relevantes. Houve tentativa de mitigar esse aspecto coletando as políticas dos aplicativos mais utilizados.

O processamento de texto tem forte dependência de ferramentas computacionais, muitas das quais são específicas para a língua portuguesa. Em especial, ferramentas para a remoção de *boilerplate* apresentaram grande variação na qualidade dos resultados. Espera-se que o crescente interesse por processamento de linguagens naturais nutra a criação e a evolução de bibliotecas para a língua portuguesa.

6. Discussões Sociológicas

A partir dos dados obtidos, verificou-se que, muitas vezes, instituições governamentais e privadas negligenciam os direitos dos cidadãos em conhecer as práticas de privacidade de dados adotadas por aplicativos para dispositivos móveis. A inacessibilidade da informação acontece quando políticas de privacidade não são disponibilizadas ou quando possuem um texto de difícil entendimento para o público comum.

Segundo a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), em 2019, 11 milhões de brasileiros com mais de 15 anos, correspondendo a 6,6% da população, eram analfabetos [IBGE 2019]. Consoante a mesma pesquisa, apenas 48,8% das pessoas com 25 anos ou mais concluíram o ensino médio. Essas taxas aumentam à medida que a faixa etária avança. A taxa de analfabetismo entre pessoas com 60 anos ou mais é de 18,0%.

Segundo a 5º edição da pesquisa Retratos da Leitura no Brasil, realizada em 2019, 48% da população brasileira não possui o hábito de leitura [IBOPE Inteligência 2019]. Tal fato, aliado à falta de educação digital da população, agrava o problema da conscientização sobre a privacidade na Internet [Soares et al. 2020].

Outro dado que deve ser considerado é o número de pessoas que possuem alguma deficiência que possa dificultar a leitura ou o entendimento de textos. Pessoas com algum grau de deficiência visual correspondem a 3,4% da população brasileira e pessoas com algum grau de deficiência mental/intelectual são 1,4% da população [IBGE 2010].

Por fim, um dado relevante, fornecido pela British Council, informa que, em 2013, apenas 5,1% da população brasileira de 16 anos ou mais possuía algum conhecimento da língua inglesa [Council 2014]. Mesmo assim, muitas políticas de privacidade são fornecidas apenas nesse idioma.

Analisando as estatísticas apresentadas, é possível concluir que textos com alto grau de dificuldade de entendimento não serão compreendidos em completo ou em partes por uma porcentagem considerável da população brasileira. Considerando que políticas de privacidade são de interesse público, é importante que elas sejam de fácil leitura pelo maior número de pessoas. Assim, é preciso que se invista mais na criação de contratos facilmente inteligíveis assim como na educação digital para a população.

7. Conclusão

Neste artigo, descreveu-se o processo de criação e de caracterização de um *corpus* de política de privacidade vigentes no Brasil. Não é de nosso conhecimento um trabalho similar a este no país.

O processo de criação envolveu o uso de ferramentas computacionais para a coleta e o processamento de dados, assim como uma revisão humana para desambiguar idiomas e identificar padrões na apresentação dos documentos.

Grande parte dos documentos coletados não consistia em políticas de privacidade ou estava escrita em inglês. Esse alto índice de inacessibilidade de informação fez com que o conjunto de documentos inicial fosse reduzido de 1.163 arquivos para 82, os quais compuseram o *corpus* final de políticas de privacidade descrito neste trabalho.

A caracterização do *corpus* envolveu duas métricas: o tamanho em número de palavras e o índice de inteligibilidade dos documentos. De forma geral, os textos das políticas são muito extensos e complexos, demandando que mesmo usuários com alto nível de instrução demorem dezenas de minutos para compreendê-los. Essa fragilidade dos contratos de privacidade vai ao encontro de trabalhos semelhantes documentados principalmente para a língua inglesa.

Pelo exposto, pode-se afirmar que muitas das políticas de privacidade com texto em português são falhas na função de informar ao público sobre práticas de uso de dados. Nesse sentido, essas conclusões contribuem para o avanço na construção de contratos digitais mais claros e objetivos.

Como trabalhos futuros, vislumbra-se uma análise semântica das políticas do *corpus* criado. Seria interessante descobrir, por exemplo, quais as informações a respeito do usuário as empresas coletam e se compartilham tais informações com terceiros. Outras preocupações estão relacionadas à anonimização de dados e ao direito ao esquecimento. Gostaríamos também de desenvolver mecanismos automatizados para as análises das políticas, em que modelos de aprendizado de máquina pudessem extrair as principais práticas de uso de dados descritas nos textos.

Agradecimentos

O presente trabalho foi realizado com apoio da Universidade Federal de Mato Grosso do Sul.

Referências

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: acesso em 17-3-2022].

Barbosa, P. H. M. (2017). Análise das permissões e violações de privacidade em aplicações para android.

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_aprendizagempesquisacompleta.pdf. [Online: acesso em 28-1-2022].

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: acesso em 28-1-2022].

IBGE (2019). PNAD Educação 2019. https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf. [Online: acesso em 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: acesso em 28-1-2022].

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.

Lei Nº 13.709 (2021). Lei geral de proteção de dados pessoais (LGPD). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. [Online: acesso em 13-12-2021].

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.

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.

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.

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)*.