
Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

# Utilizando processamento de linguagem natural para avaliar Políticas de Privacidade

Proposta de Trabalho de Graduação

**Aluno:** Rodrigo Ferreira Oliveira de Paula(rfop@cin.ufpe.br)
**Orientador:** Jéssyka Vilela (jffv@cin.ufpe.br)
**Área:** Engenharia de Software e Linguagens de Programação

Outubro/2022

Universidade Federal de Pernambuco
Centro de Informática

Rodrigo Ferreira Oliveira de Paula

# Utilizando processamento de linguagem natural para avaliar Políticas de Privacidade

*Trabalho de Conclusão de Curso apresentado no curso de Bacharelado em Ciência da computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação*

Orientadora: Jéssyka Flavyanne Ferreira Vilela

Recife

2022

# Agradecimentos

Eu quero agradecer inicialmente a minha família, minha mãe Eliete Ferreira, meu pai Joaci Albino e a minha irmã Danielly Ferreira por todo apoio ao longo do meu crescimento. Também quero agradecer minha companheira Maria Eduarda. Essa jornada pela faculdade foi cheia de desafios e se eu cheguei até aqui foi com a grande ajuda de vocês.

Meus amigos mais próximos de infância, como Rafael Tavares e João Alexandre, meus cachorros de estimação como Max, Lucky e Luna. Também qualquer amigo que também já virou família com esse tempo. Também quero agradecer a professora Jéssyka Vilela por todo apoio com esse trabalho.

*It's hard to*
*believe that it's*
*over, isn't it?*
*Funny, how we*
*get attached to*
*the **struggle***

*Old Woman - Celeste*

# Resumo

Contexto: Nos tempos atuais, com vários escândalos de empresas vazando dados de usuários em massa, existe uma insegurança dos usuários em relação às grandes empresas corporativas. Uma forma de estabelecer confiança entre empresa e usuário é a Política de Privacidade. Uma política bem escrita e obedecendo a Lei Geral de Proteção de Dados Pessoais (LGPD) no Brasil demonstra qualidade e proporciona um sentimento de segurança entre cliente e empresa. Essa segurança é importante uma vez que além de contribuir para a confiança dos usuários, atende ao direito de livre acesso já que as Políticas de Privacidade devem descrever, de forma clara e acessível, as operações executadas sobre os dados pessoais conforme requisitado pela LGPD. Problema: No entanto, verificar manualmente se uma Política de Privacidade atende a uma lei de privacidade como a LGPD, consome muito tempo e é propenso a erros uma vez que as políticas utilizam linguagens pouco claras e textos longos. Objetivo: Realizar a avaliação automatizada de Políticas de Privacidade para avaliar a qualidade de Políticas de Privacidade de empresas brasileiras, utilizando 14 critérios de avaliação de Políticas de Privacidade propostos pela literatura. Método: Para conseguir esse objetivo: (1) Foram analisadas 9 Políticas de Privacidade para identificar padrões para realizar a busca dos critérios no texto; (2) Depois, foram utilizadas técnicas de processamento de linguagem natural e criadas regras para verificar se um título é um critério válido. Também foi testado o quanto bem essa ferramenta avalia Políticas de Privacidade; e, (3) Uma avaliação da ferramenta comparando com testes manuais. Resultados: Foram analisadas 9 Políticas de Privacidade diferentes. Sobre estas políticas, a abordagem teve 5 falsos positivos. Ou seja, títulos que a ferramenta julgou como critérios mas na realidade não eram. Então a abordagem teve precisão de 93,75%. Conclusões: Devido a uma boa taxa de eficácia ficou evidente que o procedimento proposto para avaliar Políticas de Privacidade é eficiente.



Palavras-chaves: Política de Privacidade, Processamento de Linguagem Natural, Critérios de avaliação de qualidade.

# ABSTRACT

Context: In the current times, with several scandals of companies leaking user data en masse, there is an insecurity of users in relation to large corporate companies. One way to establish trust between company and user is the Privacy Policy. A well-written policy that complies with the General Law for the Protection of Personal Data (LGPD) in Brazil demonstrates quality and provides a feeling of security between customer and company. This security is important since, in addition to contributing to users' trust, it meets the right of free access, as Privacy Policies must describe, in a clear and accessible way, the operations performed on personal data as required by the LGPD. Problem: However, manually checking if a Privacy Policy meets a privacy law like LGPD is time consuming and error prone as policies use complex words and long text. Objective: To carry out automated evaluation of Privacy Policies to assess the quality of Privacy Policies of Brazilian companies, using 14 criteria for evaluating privacy policies proposed by the literature. Method: (1) Nine privacy policies were analyzed to identify patterns to perform the search criteria in the text; (2) Then, natural language processing techniques were used and rules were created to verify that a title is a valid criterion. It was also tested how well this tool evaluates privacy policies; and, (3) An evaluation of the tool against manual tests. Results: Nine different privacy policies were analyzed. On these policies, the approach had 5 false positives. In other words, titles that the tool judged being a criterion, when in reality they were not. So the approach had an accuracy of 93.75%. Conclusions: As a result of a high accuracy, it is clear that the procedure used to evaluate Privacy Policies is efficient.



**Keywords:** Privacy Policy, Privacy, Natural Language Processing, Quality Rating Criteria.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# SUMÁRIO

# 1 Introdução

## 1.1 Contexto

Hoje em dia, os dados pessoais são muito valiosos para diversos fins comerciais, incluindo propaganda [1]. Por causa disso, muitas empresas buscam dados e pagam altos valores por eles, tendo seu valor associado com o do petróleo [2]. Utilizando técnicas de estatísticas apropriadas em combinação com soluções de softwares existentes, tarefas complexas como decisões baseadas em análise de dados tornam-se mais fáceis [3].

Diante da importância do uso dos dados pessoais, surgiram leis no mundo todo para regulamentar a utilização desses dados. No Brasil, foi promulgada a Lei Geral de Proteção de Dados Pessoais (LGPD) Lei nº 13.709/2018. A LGPD possui um conjunto de direitos e obrigações para proteger os direitos fundamentais de liberdade e privacidade das pessoas que compartilham seus dados pessoais na internet [4]. Essa lei requer transparência sobre qual dado do usuário será coletado, como será utilizado e com quem será compartilhado. Também existe o direito do usuário ter fácil acesso a seus dados, poder editar ou solicitar sua exclusão. Todas essas informações devem estar disponíveis de forma transparente e acessível. Ou seja, especificadas e detalhadas na Política de Privacidade da empresa.

Nesse contexto, o processamento de dados pessoais deve ser discutido nas Políticas de Privacidade das organizações. A Política de Privacidade vem como uma ponte de comunicação entre organização e cliente para informar como ela coleta e processa os dados pessoais. A Política de Privacidade de uma empresa deve informar para que os dados são coletados e utilizados, também como são tratados, protegidos e como serão usados e transmitidos para terceiros.

Sendo assim, é importante que a política atenda a critérios que garantam a segurança, acessibilidade e direitos presentes nas leis de proteção de dados.

## 1.2 Motivação e Justificativa

As Políticas de Privacidade, em sua maioria, são difíceis de serem lidas por que possuem um vocabulário jurídico e assim de complexa interpretação [5,6]. Por exemplo, a média de legibilidade das políticas de privacidade requer ao menos dois anos de estudo superior para serem compreendidas [7]. Logo, o usuário pode possuir dificuldade de compreender o conteúdo de uma Política de Privacidade e, portanto, dificuldade de avaliar se a política atende a critérios importantes para proteger os seus dados. O trabalho investiga **qual o impacto de uma ferramenta que automaticamente analisa a qualidade de uma Política de Privacidade?**

Nesse contexto, surge a necessidade de mecanismos que auxiliem os usuários a avaliar quais informações estão sendo coletadas e como seus dados estão sendo processados. Dado o cenário apresentado, este trabalho propõe criar uma ferramenta que analisa uma Política de Privacidade por meio de critérios pré-definidos, contribuindo assim para tornar mais simples, rápido e transparente o usuário avaliar seu conteúdo.

## 1.3 Objetivos

O objetivo deste trabalho é desenvolver uma ferramenta para verificar se Políticas de Privacidade atendem a critérios de qualidade[1] propostos pela literatura.

Como objetivos específicos, propõe-se:

- Investigar técnicas de processamento de linguagem natural;
- Analisar quais critérios de avaliação de qualidade de Políticas de Privacidade podem ser avaliados de forma automatizada;
- Avaliar a eficácia da ferramenta com Políticas de Privacidade diferentes;
- Comparar as avaliações realizadas pela ferramenta com avaliações realizadas de forma manual.

[1] Disponível no Apêndice A

## 1.4 Trabalhos Relacionados

Neste trabalho, o objetivo é desenvolver uma ferramenta para automatizar a busca por critérios em uma Política de Privacidade. Nesse contexto, dois tipos de trabalhos foram analisados buscando entender o domínio do problema:

1. Critérios de avaliação de Políticas de Privacidade: o trabalho de Augusto Terra [8] foi utilizado para extrair critérios de qualidade que foram utilizados para analisar as Políticas de Privacidade pela ferramenta proposta neste trabalho.
2. Ferramentas de avaliação automatizada de políticas: trabalhos que discutem a criação de ferramentas para analisar Políticas de Privacidade.

### 1.4.1 Catálogo de critérios para avaliação de Políticas de Privacidade

Um catálogo é definido como uma lista metódica contendo informações de forma sucinta. Eles são uma síntese de várias informações de forma clara e organizada [9]. Um catálogo foi utilizado para criar critérios de avaliação de Políticas de Privacidade.

Inicialmente, Augusto Terra [8] verificou notícias sobre vazamento de dados pessoais [10] que prejudicam a sensação de segurança dos usuários em relação às empresas. Depois, foi investigado e estudado conceitos relacionados à privacidade. Um dos conceitos são as leis de privacidade. Várias leis de privacidade foram criadas para proteção dos dados pessoais dos usuários. Como, por exemplo, a LGPD [11] e *General Data Protection Regulation* [12]. Essas leis informam o conteúdo necessário para uma Política de Privacidade. Políticas de privacidade preferencialmente deveriam ser simples e organizadas. No entanto, pesquisas na literatura mostram o contrário: os documentos são longos(algumas políticas chegam a 5.000 palavras) [5] e possuem pouca clareza em alguns trechos(por exemplo, algumas falam de transferência internacional de dados pessoais, mas não falam como realizar isso)[5,6,13]. Por causa disso existe um problema: a Política de Privacidade é a

responsável por informar o usuário sobre como seus dados pessoais são gerenciados, armazenados ou compartilhados com outras empresas, porém são difíceis de compreender pelos usuários. Nesse cenário, foram criados critérios a respeito das Políticas de Privacidade para avaliar quesitos como: interpretação, compartilhamento de dados entre outros.

Utilizando uma técnica chamada *snowballing* [14] para encontrar trabalhos relevantes que fazem avaliações de Políticas de Privacidade e expõem os critérios utilizados para fazer a análise, o catálogo foi criado. Este catálogo possui 29 critérios diferentes, 12 desses critérios foram utilizados na ferramenta de automatizar a avaliação de políticas desenvolvida pelo autor deste trabalho. Esses 12 critérios foram escolhidos por serem mais fáceis de serem detectados no texto da Política de Privacidade. Outros 2 critérios foram propostos pelo autor deste trabalho. Totalizando 14 critérios[2] utilizados pela ferramenta.

### 1.4.2 Ferramentas para a análise de Políticas de Privacidade

Devido a leis de privacidade como GDPR[12] e LGPD[11], a análise de Políticas de Privacidade entrou em bastante discussão na literatura [5,6,7,15]. As leis foram criadas para proteger os dados pessoais das pessoas. Pensando nisso, vários trabalhos criaram ferramentas [16,17,18] para ajudar as pessoas a tomarem melhores decisões sobre qual empresa confiar. Cada ferramenta utilizou um conjunto de critérios para avaliar as políticas. Paradigmas como inteligência artificial, *web scraping* e outros foram utilizados de formas diferentes para chegar ao objetivo de automatizar a avaliação de uma Política de Privacidade.

O trabalho mais antigo, chamado Claudette meets GDPR [16], é sobre um projeto de pesquisa interdisciplinar feito pelo departamento de justiça da European University Institute [19]. Foi desenvolvido em cooperação entre professores da EUI com engenheiros de duas Universidades Bologna e Modena. Neste trabalho, foi observado que mesmo com o esforço da lei de privacidade (GDPR) e fiscalização do governo, muitas políticas tendem a usar palavras cláusulas injustas e ilegais[16]. Utilizando técnica de aprendizagem de máquina, chegaram à

[2] Disponível no Apêndice A

conclusão que é possível automatizar a avaliação de Política de Privacidade, se o conjunto de dados (*dataset*) for grande o suficiente.

Outro trabalho que discute o tema de automatizar avaliação de Políticas de Privacidade chamado "*An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR*" [17], observou o problema que avaliar uma Política de Privacidade manualmente consume muito tempo e também é muito propício a erros. Para ajudar a atacar esse problema, o trabalho desenvolveu o avaliador automático utilizando uma combinação de processamento de linguagem natural e aprendizado de máquina supervisionado. Primeiro, esse trabalho construiu um modelo para entender o que uma política precisa ter para atender a GDPR. Depois ele qualificou utilizando o avaliador automático se uma política em questão atende a critérios criados por esse modelo. Em seguida, utilizou 24 Políticas de Privacidade que não tinham sido avaliadas antes pela ferramenta. O resultado foi uma eficácia de 85% sobre as 24 Políticas de Privacidade não supervisionadas.

O terceiro trabalho "*A tool for analyzing privacy policies based on the LGPD*" [18], observou no Brasil uma falta de aderência à lei de privacidade (LGPD). Para auxiliar cidadãos brasileiros a proteger melhor seus dados pessoais entendendo como uma Política de Privacidade gerencia os dados, o trabalho desenvolveu um avaliador automatizado utilizando técnicas de *web scraping* e processamento de linguagem natural. O trabalho analisou 57 Políticas de Privacidade buscando informações sobre dois tópicos: coleta de dados e finalidade de tratamento dos dados. Foi desenvolvido um website em que a entrada é o endereço *url* de uma Política de Privacidade e a saída são conjuntos de palavras sobre os dois tópicos citados acima. Um ponto interessante deste trabalho foi que a ferramenta foi disponibilizada online no heroku (https://segurindo.herokuapp.com/).

Este trabalho complementa os trabalhos relacionados ao enfatizar a avaliação de políticas brasileiras em relação à LGPD e a utilização de critérios de avaliação de Políticas de Privacidade proposto no trabalho de Augusto Terra [8] a partir de estudo sistemático da literatura.

# 2. Referencial Teórico

Neste capítulo são discutidos os conceitos importantes relacionados a este trabalho: privacidade e proteção de dados pessoais, leis de privacidade, Políticas de Privacidade e critérios de avaliação de Políticas de Privacidade.

## 2.1 Privacidade e Proteção de Dados Pessoais

Devido ao crescimento de usuários na internet, criando um cenário mais dinâmico e complexo, muitos desses usuários utilizam de redes sociais para compartilhar informações. Essas informações, de forma mais específica, são dados pessoais como: nome, CPF, endereço e outros. Sobre os dados pessoais, a professora Patrícia Patrick Peck Pinheiro comenta que "*é um dos ativos mais preciosos da sociedade digital*" [20]. Devido a esse valor, existem empresas vendendo dados, grupos de hackers roubando dados, incertezas sobre como os dados são tratados, o que leva a uma necessidade de discussão sobre privacidade e proteção de dados pessoais.

No Brasil, a constituição discute sobre proteger a privacidade das pessoas [21]. Todo brasileiro tem direito à vida privada e caso ocorra uma violação deve existir uma indenização [21]. No cenário digital, o marco civil da internet inclui direitos e deveres para quem usa a internet [22]. Neste marco é incluído proteção a dados pessoais e proteção à privacidade. Porém, a lei mais importante para proteger a privacidade e dados pessoais no Brasil foi a Lei Geral de Proteção de Dados Pessoais [11]. A lei foi criada com a intenção de proteger os dados pessoais de brasileiros, de serem vendidos ou utilizados ilegalmente.

Portanto, devido ao seu valor pessoal e financeiro, é necessário um conjunto de direitos e deveres para proteger o gerenciamento desses dados pessoais. Nesse sentido, vem o conceito de leis de privacidade.

## 2.2 Leis de Privacidade

A popularização da computação pessoal foi um processo muito rápido [23]. Em poucos anos, utilizando smartphones, a maioria das pessoas consegue se comunicar o tempo todo e ficar conectado com o resto do mundo. Usamos a tecnologia para trabalhar, para se comunicar com entes queridos, para jogar jogos eletrônicos, para comprar produtos com maior facilidade e até mesmo para fazer transações financeiras [24].

Por outro lado, esse rápido crescimento no mercado de computação pessoal fez surgir diversas ameaças, e a principal delas refere-se à privacidade e proteção de dados pessoais na internet. Por esse motivo, a segurança cibernética tornou-se cada vez mais importante, e as autoridades ao redor do mundo perceberam que, sem definir leis [11,12] mais rígidas sobre tal assunto, as grandes empresas poderiam cometer abusos com dados pessoais coletados dos usuários dos serviços [25].

No Brasil, foi promulgada a Lei Geral de Proteção de Dados Pessoais (LGPD), que possui abrangência no território nacional. Outra lei de privacidade é a *General Data Protection Regulation* (GDPR). Esta, por sua vez, abrange todos os estados-membros da União Europeia. Essa lei garante que todos os cidadãos do bloco tenham o direito à proteção de dados pessoais. Essa lei foi criada com o propósito das pessoas verem os seus dados serem utilizados com responsabilidade.

Existem três pilares sobre qual a *General Data Protection Regulation* (GDPR) [26] foi criada: Governança de dados (1), Gestão de Dados (2) e Transparência de Dados (3) [27]. Governança de dados corresponde a sobre criação de uma estrutura, na qual é previamente definido quem age sobre qual informação e em qual momento, o que contribui para maior transparência e facilidade na tomada de decisões [26]. A governança de dados envolve a gestão de pessoas, tecnologia, políticas, processos envolvendo manuais de conduta de código disponíveis para todos da empresa [27].

No pilar da Gestão de Dados, existem várias regras práticas do GDPR, como a obrigação que as empresas têm de manter registros internos de todas as atividades de processamento de dados. Segundo a lei, qualquer

empresa com mais de 5 mil registros de dados em um período de 12 meses precisa ter um profissional responsável pela gestão de dados[27].

Finalmente, o pilar da Transparência de Dados aborda o consentimento do usuário. A transparência também envolve a divulgação de suas Políticas de Privacidade de maneira clara e acessível.

A empresa que não estiver em conformidade com a GDPR pode receber desde uma simples notificação (no caso de infração leve) até uma multa de €20 milhões ou de até 4% sobre a receita anual global da companhia, o que for maior [27]. Isso significa que, no caso de companhias como Google, Microsoft e Facebook, a punição pode custar bilhões de dólares.

Na américa do norte, não existe uma lei que possua abrangência nacional. Existem várias legislações sobre proteção de dados pessoais que foram elaboradas pelos estados. A mais antiga e mais conhecida é o *California Online Privacy Protection Act* (CCPA) [29] que entrou em vigor em janeiro de 2020, sendo limitada aos usuários do estado da Califórnia (EUA). O interessante é que embora sua abrangência seja bem específica, é na Califórnia que se localizam as maiores empresas de tecnologia e o Vale do Silício, logo, suas regras acabam influenciando o mercado a nível global.

Segundo o CCPA, os consumidores possuem 4 direitos [29]:

1. **O direito de ser informado**: os consumidores devem ser informados sobre os métodos de tratamento dos seus dados, antes ou durante a sua coleta.
2. **O direito de não vender seus dados**: os consumidores têm direito de escolher se seus dados pessoais podem ficar à venda ou não.
3. **O direito de não-descriminação:** um consumidor não pode sofrer descriminação por exercer seus direitos concebidos pelo CCPA.
4. **O direito de ser esquecido**: a CCPA concede aos consumidores o direito de solicitar a exclusão de quaisquer

informações pessoais coletadas em seu nome(com algumas exceções).

Os consumidores têm o direito de processar as empresas que violam a lei. As multas podem variar entre U$100 e U$750, ou qualquer valor mais alto em relação aos danos reais (no caso que estes danos podem ser comprovados). O estado pode impor multas de até U$2.500 para empresas que violarem involuntariamente a CCPA, e multas de até U$7.500 por violação, no caso de violação internacional.

No Brasil, a LGPD, que entrou em vigor em setembro de 2020, depois de vários adiamentos, posicionou o Brasil junto com outras nações que se preocupam com a proteção de dados pessoais. De modo geral, a LGPD exige que dados pessoais sejam tratados apenas para fins permitidos pela lei, específicos, explícitos e claramente definidos. Assim como o GDPR, aplicam-se também princípios de transparência, criação de processos de dados e organização de processos de dados [30]. Assim como GDPR, a LGPD aplica-se a um escopo territorial que vai além do Brasil. Isso significa estar em conformidade com ela mesmo que o negócio não esteja no Brasil. Na prática, a LGPD é aplicável se [11]:

1. As atividades de tratamento de dados da empresa forem realizadas em um servidor no Brasil;
2. A empresa oferece bens ou serviços a pessoas localizadas no Brasil;
3. A empresa trata dados provenientes de pessoas localizadas no Brasil.

Uma questão importante sobre a LGPD, são os tratamentos de dados. Os princípios são bem semelhantes aos da GDPR. A empresa só poderá utilizar os dados em pelo menos alguma dessas hipóteses [11]:

1. É necessária haver um propósito para o tratamento;
2. A maneira como os dados serão tratados e os próprios dados tratados deve estar razoavelmente alinhados ao fim do tratamento;
3. Apenas dados necessários devem ser utilizados;
4. Os usuários devem ter fácil acesso a suas informações de dados pessoais;
5. O controlador de dados deve garantir a qualidade e precisão dos dados coletados;

6. As informações e tratamento de dados devem estar facilmente disponíveis aos usuários;
7. O controlador de dados deve garantir a segurança dos dados coletados, por medidas técnicas e organizacionais;
8. Nenhum tratamento deve ser usado para fins discriminatórios;
9. O controlador de dados deve estar em conformidade com a lei e conseguir provar isso legalmente.

As corporações que violarem as diretrizes da Lei Geral de Proteção de Dados Pessoais de 2020 podem receber punições, que podem variar dependendo da gravidade da infração. As multas por não conformidade podem chegar até 2% do faturamento da organização, limitadas a R$50 milhões. Em casos mais graves, a empresa pode sofrer penalidades como ter suas atividades suspensas, parcial ou totalmente.

Após discutir sobre leis de privacidade e proteção de dados, é importante falar sobre as Políticas de Privacidade. Documentos importantes que são requeridos pelas leis de privacidade a todas as empresas.

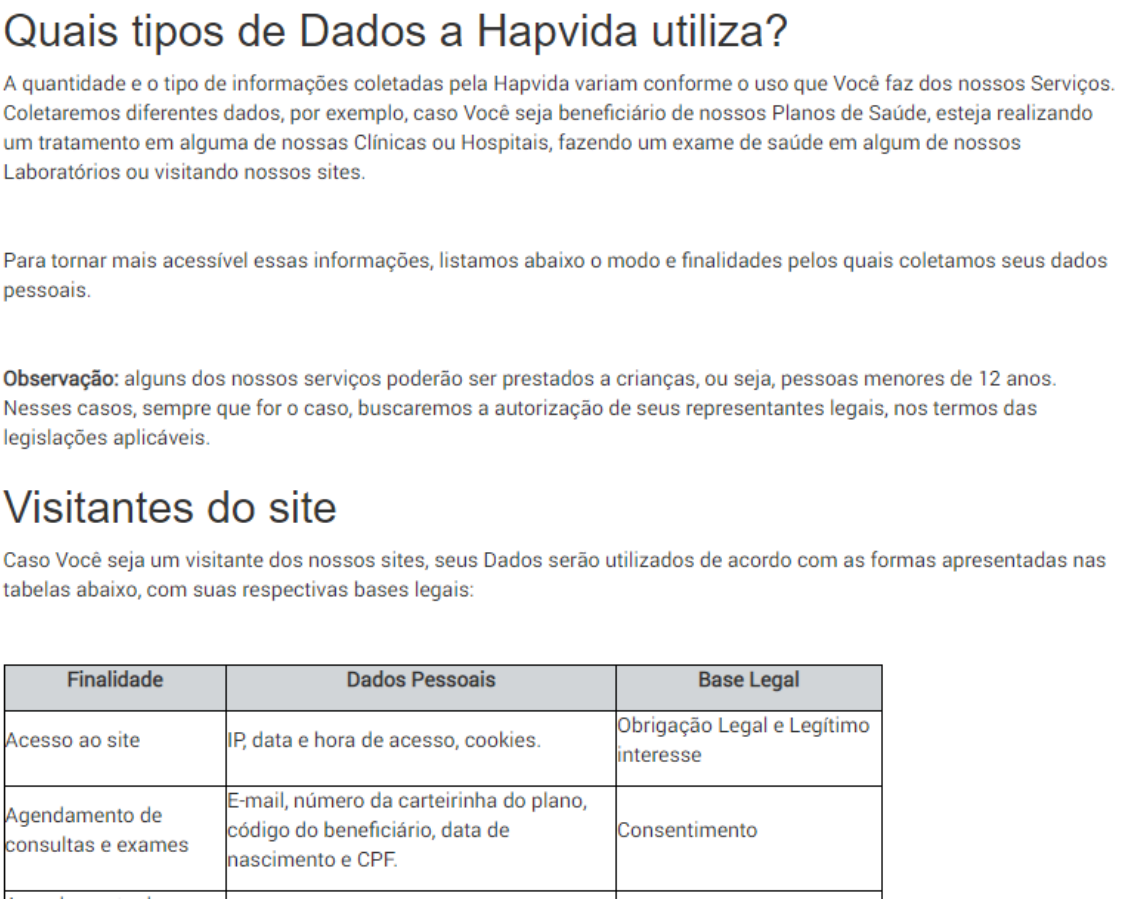
## 2.3 Políticas de Privacidade

Uma Política de Privacidade pode ser definida como um documento que explica como uma empresa gerencia os dados pessoais dos seus clientes [31]. Este documento sendo bem escrito gera um sentimento de segurança e confiança entre cliente e empresa. A política descreve quais os dados coletados, quais os dados são compartilhados, se existe transferência de dados para outra empresa, entre outras informações.

Para ilustrar, na Figura 1 é apresentado um trecho da Política de Privacidade da Hapvida. Neste trecho, é explicado como a empresa gerencia os dados pessoais. Por exemplo, uma parte discute quais dados pessoais a Hapvida utiliza. Também é descrito qual a finalidade, quais dados pessoais e a base legal. A forma como essas informações são apresentadas pode ser diversa. Algumas políticas de privacidade utilizam tabelas, imagens, vídeos e às vezes texto descritivo.

Como a lei de privacidade não impõe uma padronização no conteúdo existe toda essa diversidade na apresentação das informações. Essa falta de

padronização é um dos motivos que contribuem para a dificuldade de entender políticas de privacidade por parte dos usuários.



## Quais tipos de Dados a Hapvida utiliza?

A quantidade e o tipo de informações coletadas pela Hapvida variam conforme o uso que Você faz dos nossos Serviços. Coletaremos diferentes dados, por exemplo, caso Você seja beneficiário de nossos Planos de Saúde, esteja realizando um tratamento em alguma de nossas Clínicas ou Hospitais, fazendo um exame de saúde em algum de nossos Laboratórios ou visitando nossos sites.

Para tornar mais acessível essas informações, listamos abaixo o modo e finalidades pelos quais coletamos seus dados pessoais.

**Observação:** alguns dos nossos serviços poderão ser prestados a crianças, ou seja, pessoas menores de 12 anos. Nesses casos, sempre que for o caso, buscaremos a autorização de seus representantes legais, nos termos das legislações aplicáveis.

## Visitantes do site

Caso Você seja um visitante dos nossos sites, seus Dados serão utilizados de acordo com as formas apresentadas nas tabelas abaixo, com suas respectivas bases legais:

| Finalidade | Dados Pessoais | Base Legal |
|---|---|---|
| Acesso ao site | IP, data e hora de acesso, cookies. | Obrigação Legal e Legítimo interesse |
| Agendamento de consultas e exames | E-mail, número da carteirinha do plano, código do beneficiário, data de nascimento e CPF. | Consentimento |

Figura 1 - Trecho da Política de Privacidade da Hapvida.

Fonte: Hapvida, 2022.

No entanto, muitas Políticas de Privacidade são longas, possuem linguagem pouco clara (por exemplo declarando que “dados pessoais vão ser utilizados para melhorar o serviço” sem especificar quais dados e que tipo de melhora) e às vezes até violam a lei de privacidade vigente [5,6,12,15]. Leis de privacidade, como a LGPD, exigem que as empresas sejam transparentes para o usuário sobre as operações de tratamento de dados. Entretanto, observa-se que muitas Políticas de Privacidade não são bem escritas e não possuem boa legibilidade[5].

## 2.4. Critérios de Avaliação de Políticas de Privacidade

No empenho de auxiliar na melhoria da qualidade da escrita das Políticas de Privacidade, vários estudos na literatura foram feitos [5,6,7]. Alguns deles escolheram criar critérios de avaliação de Políticas de Privacidade [8].

Um critério pode ser definido como uma regra ou princípio para julgar, avaliar ou testar algo [32]. Alguns trabalhos desenvolveram critérios para avaliar o quanto adequada uma Política de Privacidade é a uma certa lei de privacidade vigente [8,16]. Esses critérios são utilizados para avaliar a qualidade de uma Política de Privacidade. Quanto mais critérios e quanto melhor explicado um critério em uma Política de Privacidade for, melhor será a qualidade dela. Por exemplo, um critério elaborado por Augusto Terra [8] é “A política específica claramente quais dados são coletados?”. Em uma Política de Privacidade esse critério pode ser apresentado usando uma tabela que informa os dados e seus tipos.

Dessa forma, constata-se que para proteger a privacidade e proteção de dados pessoais são necessárias leis de privacidade. A lei de privacidade impõe uma Política de Privacidade para as empresas. No entanto, elas são escritas com problemas como: sem incluir todas as informações necessárias(por exemplo falando que o titular dos dados tem direito a remover seus dados pessoais da empresa, mas não fala como fazer isso) textos longos contendo 4.191 palavras[5,6,13,16]. O que dificulta um usuário verificar se ela está de acordo com uma lei de privacidade. Para atacar esse problema, o trabalho utilizou os critérios[1] para auxiliar na verificação de qualidade de uma política de privacidade.

## 2.5 Lematização

A lematização é o processo, efetivamente, de “deflexionar” uma palavra para determinar o seu lema(as flexões chamam-se lexemas) [33]. Por exemplo, as palavras gato, gata, gatos, gatas são todas formas do mesmo lema: gato. Igualmente, as palavras “tiver”, “tenho”, "tinha", “tem” são do mesmo lema ter.

A *lematização* é útil quando queremos ver os usos de palavras em contextos sem importância das flexões [34]. Por exemplo, para a criação e uso de índices ou na investigação linguística. Assim, uma ferramenta não precisa buscar todas as formas de uma palavra para encontrá-la num texto. É por isso que a *lematização* é usada na ferramenta deste trabalho.

# 3. Metodologia

A metodologia adotada para executar esse trabalho consistiu de quatro etapas:

1. Desenvolvimento da ferramenta
2. Definição da arquitetura e tecnologias usadas
3. Elaboração do fluxo de análise das Políticas de Privacidade
4. Testes na ferramenta

## 3.1 Desenvolvimento da ferramenta

Neste trabalho foram analisadas manualmente várias Políticas de Privacidade de diferentes empresas do mercado[3]. Foi desenvolvido uma ferramenta com foco em verificar se 14 critérios[4] de avaliação de Políticas de Privacidade são contemplados em uma determinada Política de Privacidade.

Desses 14 critérios, 12 critérios foram retirados do trabalho de Augusto Terra [8] e outros 2 foram acrescentados pelo autor deste trabalho analisando as Políticas de Privacidade. O trabalho de Augusto Terra [8] propõe 29 critérios, desses constatou-se que 7[5] não eram passíveis de automatização e os demais critérios estão fora do escopo deste trabalho.

Foi realizada uma análise dos textos buscando entender a estrutura e para achar como identificar os critérios de avaliação de privacidade no documento. Depois foi feita uma seleção de palavras encontradas em títulos de Políticas de Privacidade que identificam se um critério de avaliação de

[3] Disponível no Apêndice B
[4] Disponível no Apêndice A
[5] Disponível no Apêndice C

Política de Privacidade está contemplado. A ferramenta tem duas etapas: isolar as sentenças, tratamento do texto e depois a classificação.

Primeiramente para isolar as sentenças foi realizado o seguinte:

1. O texto da Política de Privacidade é recebido por PDF como entrada.
2. Depois o texto é separado, para encontrar os títulos. Esses títulos vão ser avaliados se falam sobre critérios de avaliação de Política de Privacidade.

Tratamento das sentenças:

3. São removidos espaços vazios no começo e fim da sentenças. Todas as palavras tornam-se minúsculas para ajudar na etapa do *stop word*.
4. Como estamos buscando por títulos, parágrafos são removidos. O tamanho mínimo foi definido como sendo o maior critério encontrado nas políticas.
5. Algumas *stopwords* como a, e, o.. são removidas.
6. A lematização [33] é feita e as palavras são reduzidas, Por exemplo: dados -> dad
7. São removidos números, também são removidos palavras misturadas com números como “se7”.

Depois do tratamento no texto, inicia-se a parte de checar se as sentenças possuem as palavras-chave. Para auxiliar na validação, algumas regras[3] foram criadas para a sentença ser válida:

1. Avaliação de quais palavras[6] aparecem com mais frequência, juntando elas para compor o conjunto de palavras para um critério de avaliação da Política de Privacidade. Essas são as palavras-chave.

2. Foi encontrado o mínimo de palavras que uma sentença deve ter para ser válida para um critério de avaliação de Política de Privacidade. Alguns critérios têm apenas uma palavra no título.

3. Algumas palavras foram proibidas de estarem na sentença. Isso ajuda para um critério não ser confundido com outro.

[6] Disponível no apêndice B

4. Foi encontrado o máximo de palavras que uma sentença deve ter para ser válida para um critério de avaliação de Política de Privacidade. Isso ajuda com que um critério não seja confundido com um parágrafo.

Como resultado dessas etapas, foi criado uma ferramenta que pode ser considerada um classificador de várias classes. Cada critério possui características diferentes e essas regras[7] ajudam a auxiliar na avaliação da Política de Privacidade. A saída informa se o critério foi contemplado ou não no documento. Caso seja, o título onde o critério se encontra é informado ao usuário. Foi observado que as Políticas de Privacidade não seguem um padrão em comum. Muitas têm imagens, ilustrações o que dificulta o processamento do texto. A ordem que os critérios parecem não seguir um padrão, as palavras-chave podem diferir bastante de uma Política de Privacidade para outra. Em relação a imagens e ilustrações, a ferramenta não conseguiu extrair informações delas.

Sendo assim, a ferramenta auxilia os usuários a identificarem o quanto uma empresa contempla critérios de qualidade de Políticas de Privacidade.

Portanto, nesta seção foi explicado a forma como a ferramenta foi construída e também as etapas que o texto do documento da Política de Privacidade passa, na próxima seção são discutidas as tecnologias utilizadas.

## 3.2 Definição da arquitetura e tecnologias usadas

A ferramenta foi criada pensando na arquitetura MVC (*Model*, *View Controller*). Nesse padrão existe o *model* com seus atributos que são as regras[3] mencionadas antes, por exemplo, a quantidade mínima de palavras que uma sentença deve possuir para ser considerada válida.

Também existe um *script* que faz todo o processo de tratamento do texto e depois da validação de um critério. Por fim, as *views* são as páginas webs que o usuário interage. Foi criada uma página para o usuário inserir a Política de Privacidade no formato PDF, outra que explica um pouco sobre

[7] Disponível no Apêndice C

a ferramenta e também uma tela que explica os critérios de qualidade que são utilizados. Também existe uma tela que informa os critérios que foram encontrados ou não. Caso um critério seja encontrado, é informado o título da seção em que ele se encontra, caso não seja encontrado aparece a mensagem de “critério não encontrado”.


Figura 2 - Tela Inicial da Ferramenta. Fonte:Autor, 2022.

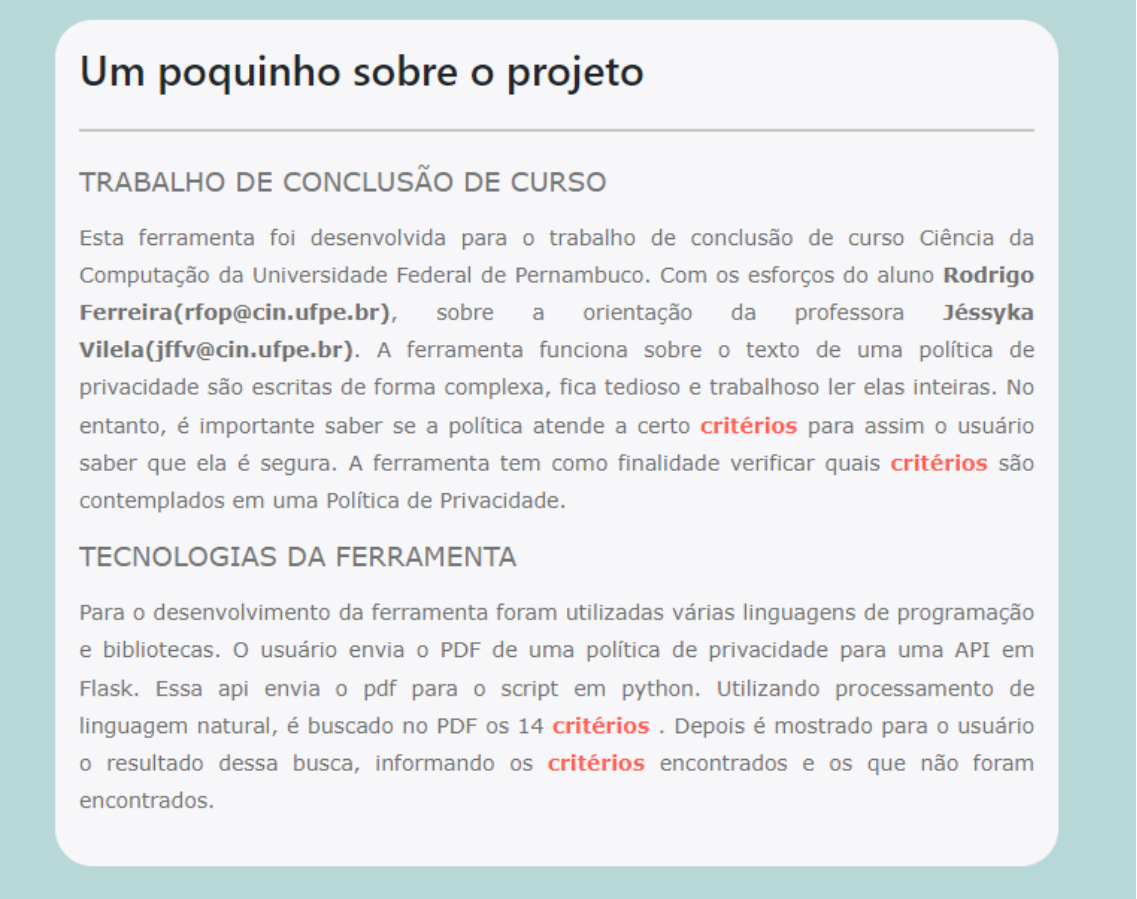
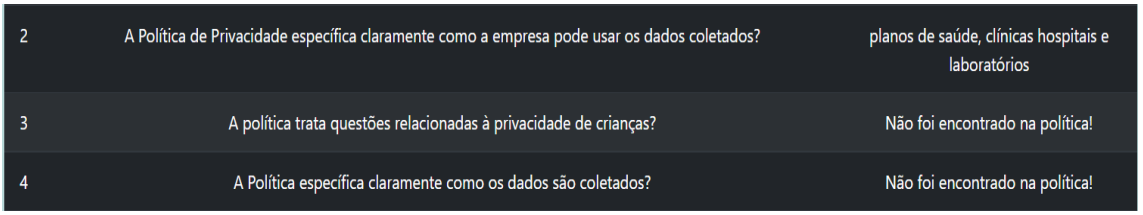



Figura 3 - Tela Sobre o Projeto. Fonte:Autor, 2022.

| # | Critério | Descrição |
|---|---|---|
| 1 | A Política específica claramente quais dados são coletados? | É importante que a Política de Privacidade detalhe claramente quais dados serão coletados pela aplicação. Os dados coletados se dividem em categorias bem definidas e a Política deve indicar essas áreas. |
| 2 | A Política de Privacidade específica claramente como a empresa pode usar os dados coletados? | A Política deve indicar qual o propósito da coleta de informações dos usuários. É necessário afirmar, por exemplo, se os dados estão sendo coletados para contactar o usuário, melhorar os serviços fornecidos, análise e monitoramento durante o uso da aplicação, personalizar a experiência, publicidade direcionada, entre outras ações. |

Figura 4 - Informação sobre os critérios utilizados pela ferramenta. Fonte:Autor, 2022.



| | | |
|---|---|---|
| 2 | A Política de Privacidade especifica claramente como a empresa pode usar os dados coletados? | planos de saúde, clínicas hospitais e laboratórios |
| 3 | A política trata questões relacionadas à privacidade de crianças? | Não foi encontrado na política! |
| 4 | A Política especifica claramente como os dados são coletados? | Não foi encontrado na política! |

Figura 5 - Tela que informa se os critérios foram encontrados ou não.
Fonte:Autor, 2022.

Para a construção da ferramenta, foi utilizado a linguagem de programação Python em conjunto com bibliotecas como NLTK, pdfminer e outras. Para fazer a requisição, o *Controller* utilizou dos protocolos de HTTP para realizar get/post de informações entre os arquivos. Para gerar as páginas webs foi utilizada a biblioteca Flask feita em Python.

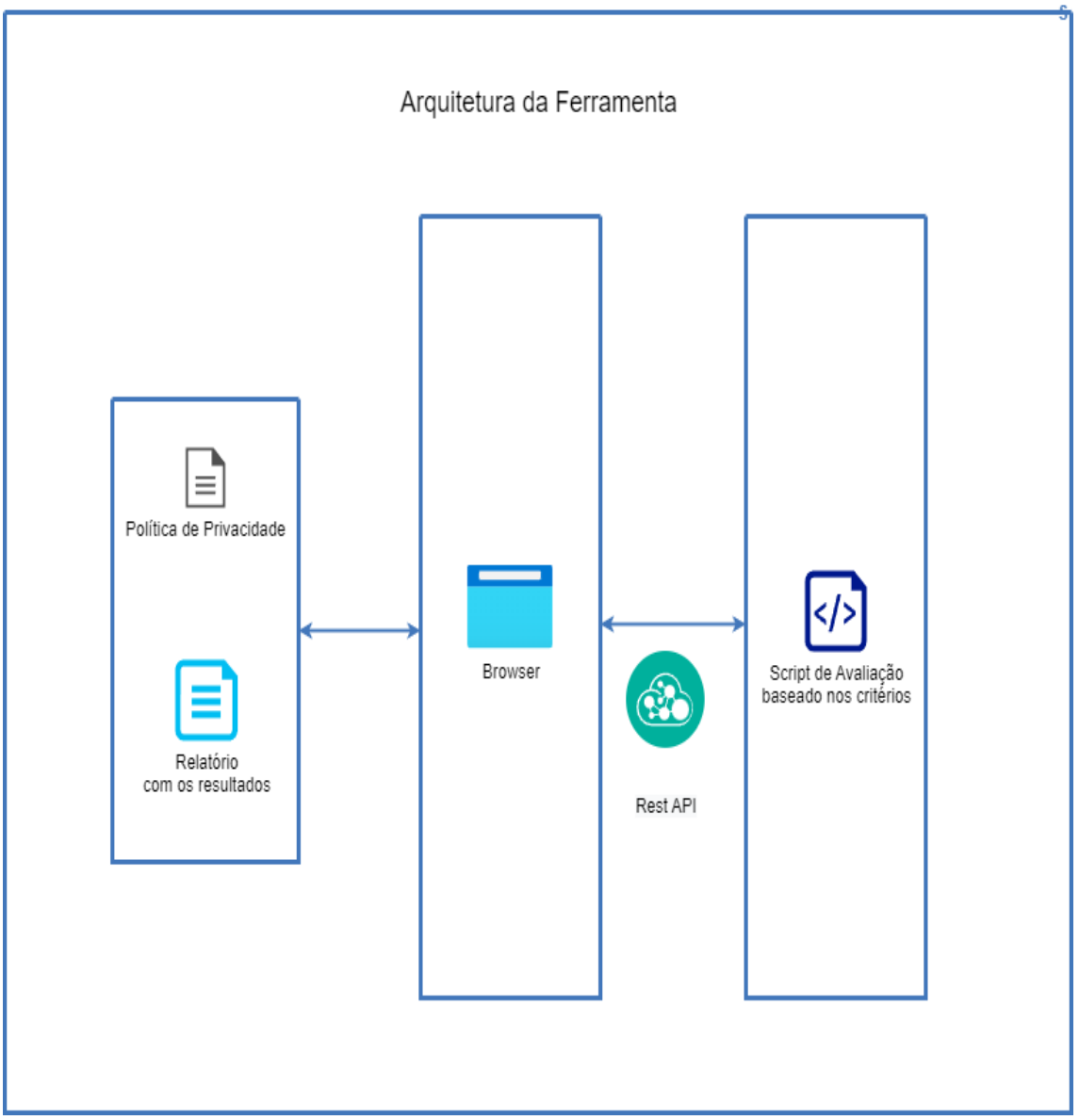



Figura 6: Diagrama conceitual da arquitetura da ferramenta.

Fonte: o autor, 2022.

## 3.3 Elaboração do fluxo de análise das Políticas de Privacidade

O objetivo da ferramenta é encontrar os critérios de qualidade nos títulos contidos em uma Política de Privacidade. Primeiro, o texto é separado utilizando regex e então o texto é filtrado para ficar apenas com títulos. Os parágrafos são descartados.

Essa busca por títulos se dá porque a ferramenta busca saber se a Política de Privacidade contempla um critério procurado. Logo, se alguma seção é válida para um critério, esse critério no mínimo é discutido no parágrafo que segue o título. Depois dos parágrafos serem filtrados, a ferramenta possui apenas um conjunto de títulos. Em seguida, é feito um filtro para remover espaços vazios, sentenças muito longas, stopwords e palavras misturadas com números. O pseudocódigo que explica esse tratamento se encontra no Algoritmo 1:

**Algoritmo 1** Tratamento do texto da política de privacidade

```
1: Lê o texto da política de privacidade
2: Percorre o texto inteiro, é obtido os títulos do texto e são apagados os parágrafos
3: Percorre as sentenças, são removidos espaços vazios antes e depois dos títulos
4: Percorre as sentenças, são removidos números e palavras misturados com números
5: Percorre as sentenças, são removidos títulos muito longos
6: Percorre as sentenças, são removidas stopwords
7: Percorre as palavras das sentenças, é feita a lematização de cada palavra
```

Figura 7: Pseudocódigo para tratamento do texto.

Fonte: o autor, 2022.

Após esse tratamento, existe a possibilidade de existir critérios de qualidade nas sentenças. Para auxiliar que um critério é válido, foram criadas algumas regras[3] descritas a seguir.

Cada critério possui diferentes regras[3] para um título ser considerado uma combinação válida. Por exemplo, para o critério “A política trata questões relacionadas à privacidade de crianças?”, as palavras chaves são 'crianç', 'adolesc', 'infantil'. Esse formato curto das palavras é efeito da lematização [33,34]. Resumidamente, a lematização é um removedor de sufixos. Esse processo auxilia em serem encontradas palavras chaves com mais facilidade. Pois, geralmente na língua portuguesa existem vários sufixos diferentes para uma única palavra.

A segunda regra é a quantidade mínima de palavras que uma sentença deve ter para ser válida. No critério sobre privacidade de crianças esse número é 1, pois existe uma política que o título é apenas “Criança”.

Essa regra ajuda a limitar o tamanho da sentença, assim evitando falsos positivos.

A terceira regra contempla sobre palavras proibidas na sentença. Essa regra foi criada para evitar que um critério seja confundido com outro pela ferramenta. No critério mencionado acima existe a palavra ‘colet’, assim evitando ser confundido com o outro critério “A política específica claramente quais dados são coletados?”.

A última regra é sobre o tamanho máximo para a sentença. Esse tamanho é definido pela maior sentença do critério encontrado nas Políticas de Privacidade avaliadas neste trabalho. Essas regras[3] são utilizadas para avaliar se um critério é válido ou não.

Para cada um dos 14 critérios[1], a função abaixo avalia se alguma sentença dentro do conjunto de sentenças é válida para ser um critério de avaliação de Política de Privacidade. Devido a falta de padronização na apresentação do conteúdo das políticas, foram propostas essas regras para auxiliar a ferramenta a encontrar os critérios.

A maior parte do desenvolvimento da ferramenta foi gasta aqui. Testes de avaliação nas regras criadas para garantir a melhor taxa de eficácia.

Por causa disso, a ferramenta diminuiu bastante a quantidade de falsos positivos. Em seguida, é apresentado os critérios encontrados e não encontrados numa tela de resultados na ferramenta. O pseudocódigo que descreve a utilização dessas regras[3] se encontra no Algoritmo 2:

**Algoritmo 2** Busca por critério válido nas sentenças

```
Essa função recebe um critério e um texto tratado pelo algoritmo 1
1: criterioLimite = criterio.qntdMinimaDePalavrasNaSenteça
2: criterioPalavraProibida = criterio.palavraProibida
3: criterionMaximoTamanho = criterio.maximoTamanho
4: Para cada sentença em textoTrado:
5:    contador = 0
6:    criterioAuxiliar = criterio.PalavrasChaves.copy()
7:    Para cada palavra em sentença:
8:          se (verifica_palavra(palavra, criterioAuxiliar) e verifica_palavraProibida(criterion
9:    Proibida, sentença):
10:               index = criterioAuxiliar.index(palavra)
11:               criterioAuxiliar[index] = ""
12:               contador = contador + 1
13:         se (contador >= criterioLimite e tamanho(sentença <= criterioMaximoTamanho)
14:               se (sentença nao está em criterion.sentenceMatch)
15:                     acrescenta a sentença em criterion.sentenceMatch
16:               acaba se
17:         acaba se
18:   acaba para cada
19:  acaba para cada
```

Figura 8: Pseudocódigo para busca por critério válido nas sentenças.

Fonte: o autor, 2022.

## 3.3 Testes na Ferramenta

Nesta seção, são apresentados os resultados de teste de eficácia com as Políticas de Privacidade. Sendo assim, é descrito como foram realizados os testes na ferramenta e seus resultados.

Primeiro vem o resultado com as políticas utilizadas na construção da ferramenta. No desenvolvimento da ferramenta, as 7 Políticas de Privacidade[8] utilizadas foram avaliadas usando os parâmetros para as regras[3]. Dos 80 títulos, 5 apresentaram falsos positivos. Isso dá uma

[8] Disponível no Apêndice B

margem de erro de 6,25% e uma taxa de eficácia de 93,75%. A ferramenta está disponível em: https://buscador-rfop.herokuapp.com/.

Para avaliar o tempo de análise das Políticas de Privacidade, foram realizadas avaliações manuais de algumas Políticas de Privacidade por estudantes de uma graduação em Computação da UFPE. Os alunos estavam matriculados em uma disciplina cujo foco principal era desenvolvimento de software em ambientes regulados e privacidade. Os alunos tinham familiaridade com as Leis de Privacidade GDPR e LGPD bem como Políticas de Privacidade. A média de idade dos alunos é 22 anos.

Seis alunos divididos em 3 duplas avaliaram políticas utilizadas na construção da ferramenta e também outras novas conforme apresentado na Tabela 1.

Tabela 1. Perfil dos alunos.

| **Dupla** | **Políticas usadas na construção da ferramenta que foram avaliadas** | **Política escolhida pela dupla para avaliação** |
|---|---|---|
| Dupla 1 | Kabum, Spotify | Steam |
| Dupla 2 | FastShop, Linkedin | Nestlé |
| Dupla 3 | Hapvida, Twitter | Crunchyroll |

Para avaliar cada política, eles demoraram em torno de 30 minutos, alguns alunos utilizaram o critério de pesquisa do navegador para análise ser mais rápida, reduzindo o tempo de análise para 20 minutos. Por outro lado, a ferramenta gasta 3 segundos para realizar a análise.

Para políticas avaliadas durante a construção da ferramenta a taxa de eficácia dos estudantes foi 75%, em paralelo, a da ferramenta foi de 93,75%. Para políticas não avaliadas durante a construção da ferramenta, a taxa de eficácia dos estudantes foi quase a mesma 70%. A ferramenta, porém, diminuiu para 48%. Isso aconteceu devido à diversidade em que as informações são apresentadas. Devido à utilização de tabelas, ilustrações para exibir os dados. Às vezes, o documento não é lido devido ao seu tipo. A tabela 2 ilustra as informações observadas com os testes na ferramenta.

Tabela 2. Informações observadas com testes na ferramenta

| | Tempo de análise | Eficácia da análise das políticas usadas para treinar a ferramenta | Eficácia da análise das políticas novas |
|---|---|---|---|
| Análise manual dos alunos | Em torno de 30 minutos, alguns alunos utilizaram o critério de pesquisa do navegador para análise ser mais rápida, reduzindo o tempo de análise para 20 minutos. | 75% | 70% |
| Ferramenta | 3 segundos | 93,75% | 48% |

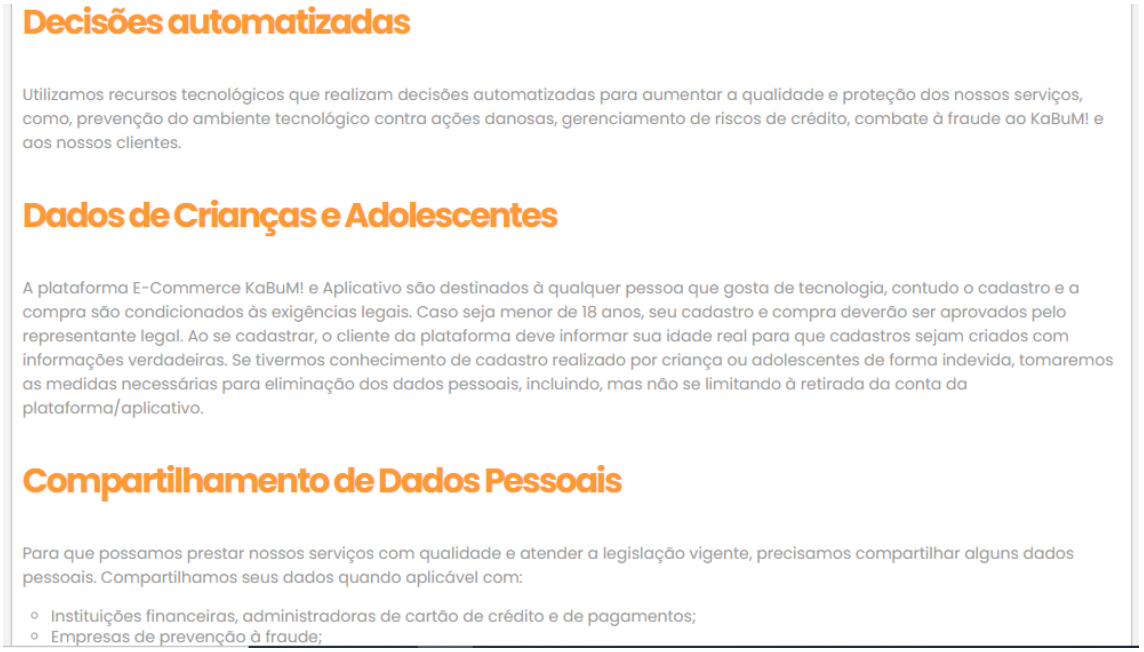
As limitações podem ser resolvidas utilizando mais políticas no momento de desenvolvimento da ferramenta e também na utilização de aprendizagem de máquina para avaliar palavras.

## 4. Resultados e Discussão

Nesta seção, são apresentados os resultados da ferramenta passo a passo para uma Política de Privacidade. Em seguida, são apresentados comparações com outros trabalhos, limitações da ferramenta e sugestões para trabalhos futuros.

## 4.1 Exemplo de utilização da ferramenta

Para exemplificar a utilização da ferramenta foi escolhida a Política de Privacidade da Kabum (https://www.kabum.com.br/privacidade). A figura abaixo mostra um trecho da Política de Privacidade:



## Decisões automatizadas

Utilizamos recursos tecnológicos que realizam decisões automatizadas para aumentar a qualidade e proteção dos nossos serviços, como, prevenção do ambiente tecnológico contra ações danosas, gerenciamento de riscos de crédito, combate à fraude ao KaBuM! e aos nossos clientes.

## Dados de Crianças e Adolescentes

A plataforma E-Commerce KaBuM! e Aplicativo são destinados à qualquer pessoa que gosta de tecnologia, contudo o cadastro e a compra são condicionados às exigências legais. Caso seja menor de 18 anos, seu cadastro e compra deverão ser aprovados pelo representante legal. Ao se cadastrar, o cliente da plataforma deve informar sua idade real para que cadastros sejam criados com informações verdadeiras. Se tivermos conhecimento de cadastro realizado por criança ou adolescentes de forma indevida, tomaremos as medidas necessárias para eliminação dos dados pessoais, incluindo, mas não se limitando à retirada da conta da plataforma/aplicativo.

## Compartilhamento de Dados Pessoais

Para que possamos prestar nossos serviços com qualidade e atender a legislação vigente, precisamos compartilhar alguns dados pessoais. Compartilhamos seus dados quando aplicável com:

- Instituições financeiras, administradoras de cartão de crédito e de pagamentos;
- Empresas de prevenção à fraude;

Figura 9: Trecho da Política de Privacidade da Kabbum. Fonte: Kabum, 2022.

Com o link da Política de Privacidade o usuário baixa a política no formato PDF. Depois o usuário acessa a ferramenta pelo *link* https://buscador-rfop.herokuapp.com/. Depois de selecionar o PDF baixado no seu computador, basta apenas clicar em buscar critérios. A figura abaixo mostra a tela da ferramenta:

Figura 10: Tela inicial da ferramenta com PDF selecionado. Fonte: o autor, 2022.

Então o usuário é redirecionado para a tela de apresentação de resultados, exibindo os critérios encontrados e os não encontrados. A figura abaixo mostra o resultado:

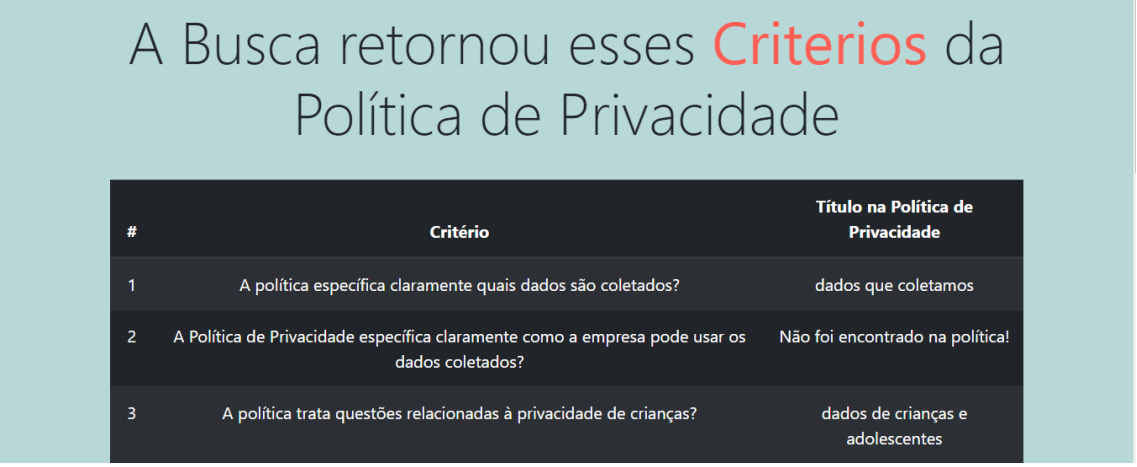



Figura 11: Tela de resultados da ferramenta com PDF selecionado. Fonte: o autor, 2022.

## 4.2 Comparação das Abordagens

A partir da comparação dos trabalhos relacionados na Tabela 3, observou que existe na literatura estudos para automatizar as Políticas de Privacidade usando várias técnicas diferentes.

Tabela 3. Comparação entre o trabalho proposto e trabalhos relacionados.

| **Critério** | **CLAUDETTE meets GDPR** | **An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR** | **A tool for analyzing privacy policies based on the LGPD** | **Este trabalho** |
|---|---|---|---|---|
| **Ano** | 2018 | 2020 | 2022 | 2022 |
| **Quantidade de políticas avaliadas** | 14 | 234 | 57 | 7 |

| | | | | |
|---|---|---|---|---|
| **Domínio dos sites** | Grandes empresas de tecnologia do mundo | Empresas da indústria de fundo monetário | Empresas brasileiras de comunicação, do governo brasileiro e de notícias diversas | Empresas de diversas áreas do mercado |
| **Paradigmas utilizados** | Inteligência artificial(aprendizagem de máquina) | Processamento de linguagem natural e aprendizagem de máquina supervisionado | Web scraping, REST API, Vue e processamento de linguagem natural (NLTK) | Processamento de linguagem natural(NLTK), Rest API e Flask |
| **Lei de privacidade** | GDPR | GDPR | LGPD | LGPD |
| **Porte das empresas avaliadas** | Grandes | Grandes | Grandes e médias | Grandes e médias |
| **Critérios para a avaliação** | Clareza da informação, conformidade com o processamento de dados da GDPR e abrangência da informação. Esses três critérios são divididos em outros menores. | Processamento de dados pessoais, conformidade com a lei vigente e outros critérios. | Sobre dados coletados e sobre finalidade de tratamento | A Política específica claramente quais dados são coletados?<br>A Política de Privacidade específica claramente como a empresa pode usar os dados coletados?.. e mais. Todos os critérios utilizados estão na tabela: lista de critérios de avaliação de Políticas de Privacidade |
| **Quantidade de critérios avaliados** | 3 | 47 | 2 | 14 |
| **Taxa de eficácia** | 75% | 85% | Não informado | 93,75% |

É interessante observar que todos os trabalhos verificaram que nas políticas faltam informações(por exemplo se um determinado dado pessoal que não era coletado, passar a ser e às vezes até colocando cláusulas ilegais conforme a lei de privacidade regente [16]. Todas utilizaram processamento de linguagem natural e algumas aprendizagem de máquina. O trabalho Claudette meets GDPR[16] possui uma análise jurídica e de software. Pois

foi uma junção multidisciplinar de advogados e engenheiros. Eles fizeram uma análise da clareza das palavras, eficácia da ferramenta e se existiam cláusulas que quebram a conformidade com a GDPR. Em comparação o trabalho A tool for analyzing privacy policies based on the LGPD focou em analisar se as palavras encontradas no conteúdo eram relacionados a dois critérios(sobre dados coletados e sobre finalidade de tratamento), testes de integração e usabilidade(por exemplo opção de libras para os usuários da ferramenta que precisam).

# 5. Conclusão

Neste trabalho, foi proposta uma ferramenta utilizando processamento de linguagem natural para avaliar a qualidade de Políticas de Privacidade por meio de critérios de avaliação de Políticas de Privacidade propostos no catálogo de Augusto Terra [8].

A justificativa para o desenvolvimento da ferramenta é que avaliar manualmente uma política de privacidade é demorado e propenso a erros. Isto ocorre por que geralmente as políticas de privacidade são escritas com pouca clareza, confundindo assim os leitores na hora da avaliação. Por exemplo, declarando que “dados pessoais vão ser utilizados para melhorar o serviço” sem especificar quais dados pessoais e que tipo de melhora. Outro problema comum é que elas são longas. Mais especificamente a quantidade de palavras numa política de privacidade pode conter 4.191 palavras ou mais, sendo assim tendo número de palavras próxima a um livro[35].

Nesse contexto, foi construída uma ferramenta automatizada para classificar o conteúdo de uma Política de Privacidade e assim verificar a qualidade da mesma usando 14 critérios. A ferramenta foi disponibilizada como página *web* para os usuários avaliarem a política que desejarem. Assim sendo de fácil acesso a usuários da internet.

Ao decorrer do desenvolvimento, foram encontrados problemas devido à falta de padronização da escrita das Políticas de Privacidade. Também existe uma dificuldade em formar o conjunto de palavras necessárias para classificar um critério devido à língua portuguesa ser complexa, por que existe a possibilidade de uma mesma palavra ter significados diferentes. Isso resultou em limitações na ferramenta por que resulta em critérios não encontrados ou falsos positivos.

Como trabalhos futuros, observa-se as seguintes oportunidades de pesquisa:

- buscar soluções para os desafios da língua portuguesa;
- realizar teste de usabilidade na ferramenta;

- definição de um formato padrão para os documentos das Políticas de Privacidade;
- implementar aprendizagem de máquina ou técnicas mais robustas de processamento de linguagem natural como *word embedding*.

# Referências


[1] De Lima, C, E 2018. Os dados como base para criação de um método de planejamento de propaganda

[2] Berardi, Rita Cristina Galarraga. Dados: o mais valioso recurso no mundo atual?. Perspectivas do Discurso Jurídico, p. 9.

[3] Santos, Fernando António Castilho Mamede dos; Santos, Fernando Miguel Soares Mamede dos. O marketing e a análise de dados para a tomada de decisões. Millenium, p. 168-177, 2004.

[4] Lei Geral de Proteção de dados Disponível em: https://www.gov.br/cidadania/pt-br/acesso-a-informacao/lgpd Acesso em 07/10/2022

[5] - Singh, Ravi Inder, Manasa Sumeeth, and James Miller. "Evaluating the readability of privacy policies in mobile environments." International Journal of Mobile Human Computer Interaction (IJMHCI) 3.1 (2011): 55-78.

[6] - Jenssen, Pots - “Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices”

[7] ] - J. Graber, D. D’Alessandro, and J. Johnson-West, “Reading level of privacy policies on Internet health websites,” J. Family Practice, vol. 51, no. 7, pp. 642- 645, 2002.

[8] Terra, Augusto Henriques. Catálogo de critérios para avaliação de políticas de privacidade. UNIVERSIDADE FEDERAL DE PERNAMBUCO, 2021.

[9]- Definição - Catálogo Dísponivel em : https://dicionario.priberam.org/cat%C3%A1logo. Acesso em 07/10/2022

[10] Uber admite ter escondido vazamento de dados 57 milhões de usuários. Disponível em: https://www.tecmundo.com.br/seguranca/242206-uber-admite-ter-escondido-vazamento-dados-57-milhoes-de-usuarios.htm. Acesso em 07/10/2022

[11] - 2019. BRASIL. Lei Geral de Proteção de Dados Pessoais (LGPD) Ministério da Cidadania,https://www.gov.br/cidadania/pt-br/acesso-a-informacao/lgpd/lei-geral-de-protecao-de-dados-pessoais-lgpd. Acesso em 07/10/2022.

[12] GDPR - https://gdpr.eu/. Acesso em 07/10/2022.

[13] - Miller, J, 2016. Evaluating the Readability of Privacy Policies
[14] - Claes Wohlin - Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering

[15] Egberto S. Diretrizes para apresentação de Políticas de Privacidade voltadas para a Experiência do Usuário.
[16] - Giuseppe Contissa - Koen Docter - Francesca Lagioia - Marco Lippi Hans-W. Micklitz - Przemyslaw Palka - Giovanni Sartor - Paolo Torroni - CLAUDETTE meets GDPR
[17] - Damiano Torrea , Sallam Abualhaijaa , Mehrdad Sabetzadehb, Lionel Brianda Katrien Baetens, Peter Goes, and Sylvie Forastier - An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR
[18] - Victor Matheus Pereira de Azevedo e Victor Rodrigues Marques - A tool for analyzing privacy policies based on the LGPD
[19] EUI - https://www.eui.eu/en/home. Acesso em 07/10/2022.
[20] Alfredo de J. Flores, Alejandro Montiel Alvarez, Anderson Vichinkeski Teixeira, Wagner Feloniuk. Dados: o mais valioso recurso no mundo atual?
[21] - BRASIL. Constituição (1988). Constituição da República Federativa do Brasil.
[22] -BRASIL. MARCO CIVIL (2015) https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/direito-facil/edicao-semanal/marco-civil-da-internet#:~:text=O%20Marco%20Civil%20da%20Internet,da%20internet%20no%20Brasil…
[23] Windows que ajudou a popularizar os computadores pessoais, faz 30 anos. Disponível em: https://m.folha.uol.com.br/tec/2015/11/1708844-windows-que-ajudou-a-popularizar-os-computadores-pessoais-faz-30-anos.shtml. Acesso em 07/10/2022
[24] Entenda a revolução promovida pelo pix no Brasil; Disponível em: https://publicacoes.estadao.com.br/financasmais/2021/08/17/entenda-a-revolucao-promovida-pelo-pix-no-brasil/#:~:text=Desde%20sua%20estreia%20em%20novembro,revolu%C3%A7%C3%A3o%20na%20vida%20dos%20brasileiros. Acesso em 07/10/2022

[25] Facebook compartilhou mais dados com gigantes tecnologicos do que o revelado. Disponível em https://g1.globo.com/economia/tecnologia/noticia/2018/12/19/facebook-compartilhou-mais-dados-com-gigantes-tecnologicos-do-que-o-revelado-diz-jornal.ghtml. Acesso em 07/10/2022

[26] Gawronski, M. (2019). Guide to the GDPR. Kluwer Law International BV.

[27] Isabel M. "UMA PROPOSTA DE GOVERNANÇA DE DADOS BASEADA EM UM MÉTODO DE DESENVOLVIMENTO DE ARQUITETURA EMPRESARIAL"

[28] NAM, Yeonghun; SHIN, Eunok; SUYEONG, Lee; SEUNGHO, Jung; BAE, Yohan; JUNGHYUN, Kim. Global-scale GDPR Compliant Data Sharing System. 2020 International Conference on Electronics, Information, and Communication (ICEIC)

[29] - CCPA - California Consumer Privacy Act (CCPA) https://oag.ca.gov/privacy/ccpa. Acesso em 07/10/2022

[30] Evandro T. Transações de Dados e Privacidade à luz da Lei Geral de Proteção de Dados Pessoais (LGPD)

[31] - Definição Política de Privacidade. Disponível em: https://www.techtarget.com/whatis/definition/privacy-policy. Acesso em 07/10/2022.

[32] - Definição de critério. Disponível em: https://www.dictionary.com/browse/criterion. Acesso em 07/10/2022

[33]De Lucca, J. L., & Nunes, M. D. G. V. (2002). Lematização versus Stemming. USP, UFSCar, UNESP, São Carlos, São Paulo.

[34] - Projeto de Lematização . Disponível em: https://www.inf.ufrgs.br/~viviane/rslp/index.htm. Acesso em 07/10/2022

[35] - -We Read 150 Privacy Policies. They were an Incomprehensible Disaster. Disponível em: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html. Acesso em 07/10/2022

# Apêndice A - Lista de Critérios de Avaliação de Política de Privacidade

| Critério | Descrição |
|---|---|
| A política específica claramente quais dados são coletados? | É importante que a Política de Privacidade detalhe claramente quais dados serão coletados pela aplicação. Os dados coletados se dividem em categorias bem definidas e a Política deve indicar essas áreas. |
| A Política de Privacidade específica claramente como a empresa pode usar os dados coletados? | A Política deve indicar qual o propósito da coleta de informações dos usuários. É necessário afirmar, por exemplo, se os dados estão sendo coletados para contactar o usuário, melhorar os serviços fornecidos, análise e monitoramento durante o uso da aplicação, personalizar a experiência, publicidade direcionada, entre outras ações. |
| A política trata questões relacionadas à privacidade de crianças? | É necessário que a Política explique claramente como se dá questões relacionadas à privacidade com crianças que acessam a aplicação. |
| A Política específica claramente como os dados são coletados? | A Política precisa expressar com clareza quais ferramentas a aplicação utiliza para coletar dados. |
| A Política de Privacidade claramente especifica se as informações podem ser compartilhadas ou vendidas para terceiros? | Caso envolva terceiros, é necessário descrever que tipo de informações são compartilhadas, quem são os terceiros e como os terceiros podem ser classificados, além de estar anexada a Política de Privacidade dessa empresa terceira. É necessário afirmar também caso não haja o compartilhamento com outras organizações. |
| Decisões Automatizadas | Aqui o critério verifica se a política discute se existem recursos tecnlógicos que realizam decisões automatizadas para fim de melhorar o serviço prestado pela empresa |
| A Política de Privacidade claramente especifica quais são as medidas adotadas pela aplicação para garantir a confidencialidade, a | Este critério busca avaliar se a aplicação possui algum método para garantir a confidencialidade e integridade dos dados do |

| | |
|---|---|
| integridade e a qualidade dos dados? | usuário. Por exemplo, se o armazenamento dos dados é criptografado ou alguma máscara de IP é utilizada. |
| A política explica claramente o que acontece com os dados do usuário caso ele exclua a conta? | É importante que esteja descrito na política o que acontece caso o usuário se desvincule da aplicação. |
| A Política de Privacidade claramente especifica os direitos do usuário? | As leis de privacidade apresentam direitos que os usuários possuem. É uma boa prática que a política descreva esses direitos em relação a seus dados pessoais. |
| A Política de Privacidade fala sobre como ela utiliza cookie no seu site? | Este critério busca avaliar se o site fala sobre os tipos de cookies utilizando pelo website |
| A Política de Privacidade claramente informa dados para contato com a empresa? | Idealmente deve haver o contato da área da empresa que trate de questões de privacidade dos dados de seus usuários. |
| A Política de Privacidade claramente específica como os dados são armazenados? | Ao informar como os dados são armazenados a empresa passa uma maior credibilidade para seus usuários. |
| A Política de Privacidade fala sobre transferir dados do usuário em nível internacional? | O critério idealmente deve falar sobre como transferir os dados do cliente para outras regiões fora do Brasil. |
| Como as alterações nas políticas são tratadas? | Após uma eventual alteração na Política de Privacidade, os usuários precisam ser informados e notificados sobre isso. |

## Apêndice B - Lista de Políticas de Privacidade

| Empresa | Link da política | Utilizada no desenvolvimento? |
|---|---|---|
| Epic games | https://www.epicgames.com/site/pt-BR/privacypolicy?sessionInvalidated=true | Sim |
| Fast Shop | https://empresas.fastshop.com.br/politica-de-privacidade/content/16#:~:text=Esta%20Pol%C3%ADtica%20disp%C3%B5e%20sobre%20o,escrit%C3%B3rios%20na%20condi%C3%A7%C3%A3o%20de%20cliente%2C | Sim |
| Hapvida | https://www.hapvida.com.br/site/politicas-de-privacidade | Sim |
| Kabbum | https://www.kabum.com.br/privacidade?gclid=Cj0KCQjwkOqZBhDNARIsAACsbfLRpuPTuA50HHDHqDkzqTX3kvLsfIzjCpQd0-ZKyMiTi-Oz2l3Z4IQaAiPtEALw_wcB | Sim |
| Linkedin | https://br.linkedin.com/legal/privacy-policy | Sim |

| | | |
|---|---|---|
| Spotify | https://www.spotify.com/br/legal/privacy-policy/ | Sim |
| Steam | https://store.steampowered.com/privacy_agreement/brazilian/?l=portuguese | Sim |
| Nestlé | https://www.nestle.com.br/politica-de-privacidade | Não |
| Crunchyroll | https://www.crunchyroll.com/pt-br/privacy | Não |

# Apêndice C - Lista Critérios Não Possíveis de Automatizar

| Critérios | Descrição |
|---|---|
| A aplicação apresenta a Política de Privacidade no momento que o usuário acessa a plataforma? | Para esse critério é necessário avaliar se o usuário consegue ter acesso à Política de Privacidade por meio de um link externo ou um pop-up assim que ele entra na aplicação. |
| Quão fácil é para o usuário encontrar a Política de Privacidade na aplicação? | Local onde o link para a Política de Privacidade está alocado e qual a visibilidade dele para o usuário. |
| O documento está devidamente traduzido para todas as línguas que a aplicação suporta? | É necessário que a Política de Privacidade da aplicação esteja corretamente escrita e traduzida para todos os idiomas que a aplicação dá suporte. Isso vai garantir que o documento passe credibilidade para o usuário confiar suas informações pessoais à organização. |
| A Política de Privacidade é acessível para Pessoas Com Deficiência (PCD)? | O texto do documento deve apresentar ajustes visuais, como aumento e diminuição da fonte do texto ou ajuste de cores, para auxiliar a leitura de usuários com baixa visão. Outra opção é incluir leitores de tela automatizados, por meio de voz sintetizada, que leem o conteúdo da Política e expõem para o usuário. |
| O documento é responsivo? | É necessário que a Política esteja disponível em dispositivos com diferentes tamanhos de tela, seja desktop ou mobile. |
| O documento apresenta uma boa usabilidade em dispositivos com diferentes tamanhos de tela? | A experiência do usuário ao ler a Política de Privacidade deve ser boa, inclusive em dispositivos móveis. É preciso ter cuidado com o tamanho da fonte e tamanho do documento. Como muitos usuários acessam aplicações através do celular deve-se levar em consideração esse tamanho de tela para a escrita da política. |

| | |
|---|---|
| A política menciona sobre o acesso de pessoas menores de idade? | Caso a aplicação permita o acesso a pessoas menores de idade, é preciso que a política de privacidade aborda esse tema. |

## Apêndice D - Lista de Regras para Cada Critério

| Nome do critério | Palavras chaves | Quantidade mínima de palavras numa sentença para a sentença ser considerada válida | Palavras Proibidas | Quantidade máxima de palavras que uma sentença pode ter para ser considerada válida |
|---|---|---|---|---|
| A política específica claramente quais dados são coletados? | 'dad','colet','tip','qual', 'que','guard', 'registr', 'ativ', 'inform','de', 'visit', 'sit','do' | 3 | 'motivo', 'proteg', 'cooki','client', 'tecnolog', 'automatic', 'direit', 'requis','assess', 'marketing','proteç','oferec','produz','realiz','endereç', 'plataform', 'login', 'log','final', 'naveg','funcional','internet', 'consent', 'além','categor', 'terc','permiss','permit','lei','c','seguint','país','reg','serviç','app','seç' | 9 |
| A Política de Privacidade específica claramente como a empresa pode usar os dados coletados? | us','obje', 'dad', 'recolh', 'motiv', 'final', 'pessoal', 'uso', 'com', 'inform', 'plan', 'saúd', 'de','clín','hospit','laboratóri | 3 | 'disposi','trat','sobr', 'compartilh', 'direit', 'requis','proteg', 'receb','descobr','parc','ferrame | 6 |

| | | | | |
|---|---|---|---|---|
| | | | nt','colet','seguint','tip','cri','produz','exclus','pag','gent','outr' | |
| A política trata questões relacionadas à privacidade de crianças? | 'crianç','adolesc', 'infantil' | 1 | 'destin','colet','cont' | 8 |
| A Política específica claramente como os dados são coletados? | 'dad', 'colet', 'com', 'inform' | 3 | 'operac', 'ferrament', 'interrupç','opç','seguint','disposi','reg' | 8 |
| A Política de Privacidade claramente especifica se as informações podem ser compartilhadas ou vendidas para terceiros? | 'dad', 'compartilh', 'divulg', 'pessoal', 'com', 'acess', 'qu', 'inform' | 3 | 'polít', 'plataform', 'sign', 'aplic', 'seguint', 'requis','proteg','receb','solic','descobr','escolh','direit','exclu','divers','spot','promov','colig', 'segurançç','reg', 'trat','outr','sab', 'atual','disposi' | 8 |
| Decisões Automatizadas | decis', 'autom' | 2 | | 2 |
| A Política de Privacidade claramente especifica quais são as medidas adotadas pela aplicação para garantir a confidencialidade, a integridade e a qualidade dos | 'segurançç', 'compromet', 'proteg' | 1 | 'nacion','alguém', 'internet', 'códig', 'assegur', 'temp', 'cont','nossos' | 5 |

| | | | | |
|---|---|---|---|---|
| dados? | | | | |
| A política explica claramente o que acontece com os dados do usuário caso ele exclua a conta? | 'dad', 'elimin', 'exclus', 'inform' | 2 | 'final','receb','identific','colet','gent' | 6 |
| A Política de Privacidade claramente especifica os direitos do usuário? | 'direit', 'titul', 'control', 'opç', 'seu','obrig' | 2 | 'exerç' | 6 |
| A Política de Privacidade fala sobre como ela utiliza cookie no seu site? | 'cooki', 'polít', 'colet','utiliz','com','de' | 3 | 'tip','gerenci','colet' | 6 |
| A Política de Privacidade claramente informa dados para contato com a empresa? | 'inform','contat', 'contact', 'conosc', 'fal', 'gent', 'com','hapv' | 2 | armazen','us','compartilh','segurança', 'proteg', 'colet','especí','receb','content','reclam', 'calend','ped','pessoal','cont','autoridad','sobr', 'pag' | 5 |
| A Política de Privacidade claramente específica como os dados são armazenados? | 'armazen', 'retenç', 'temp','dur','dad' | 2 | 'país','ambi','cadastr','própri','preserv','qualqu','sistem' | 8 |
| A Política de Privacidade fala sobre transferir | 'transfer','internac','país', 'glob', 'oper','inform','califórn','europ' | 2 | ferrament', 'control' | 8 |

| | | | | |
|---|---|---|---|---|
| dados do usuário em nível internacional? | ,'fronteir' | | | |
| Como as alterações nas políticas são tratadas? | 'atual','alter', 'mudanç' | 1 | 'cadastr','novembr','hábit','control','kabum','ped' | 5 |