Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explorando Princípios da General Data Protection Regulation, Lei Geral de Proteção de Dados e Diretrizes Éticas da Inteligência Artificial em Repositórios Open Source

Bernardo Ferreira Santos Ximbre

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2022

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explorando Princípios da General Data Protection Regulation, Lei Geral de Proteção de Dados e Diretrizes Éticas da Inteligência Artificial em Repositórios Open Source

Bernardo Ferreira Santos Ximbre

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof.a Dr.a Leticia Lopes Leite Prof.a Dr.a Fernanda Lima

Prof. Dr. João José Costa Gondim
Coordenador do Curso de Engenharia da Computação

Brasília, 27 de abril de 2022

Dedicatória

Eu dedico este trabalho aos meus pais, Maria Helena e Mauri que me acompanharam em toda essa caminhada acadêmica e também a professora Edna Dias Canedo que ajudou a elaborar e a finalizar este ciclo em minha vida.

Agradecimentos

Agradeço a todos que fizeram parte dessa caminhada de anos que me fez pensar diversas vezes em desistir. Agradeço a minha família que sempre buscou apoiar nesses momentos difíceis da jornada. Agradeço ainda ao meu ciclo de amigos que obtive ao longo dessa caminhada, fazendo com que todos se ajudassem e chegassem ao final. Por fim agradeço a todas as pessoas que de certa forma me ajudaram indiretamente mesmo sem me conhecer, servidores da limpeza, servidores do restaurante universitário, do transporte público, onde foram responsáveis por prestar serviços essenciais como deslocamento, alimentação e conforto para o meu dia a dia dentro e fora do ambiente acadêmico.

Resumo

Com o desenvolvimento da tecnologia os dados dos usuários se tornaram um dos bens mais importantes para as empresas, seguido disso o surgimento e uso descontrolado da Inteligência Artificial (IA) colocou em risco a privacidade desses usuários. Dado este cenário, surgiu a necessidade de criar uma regulação que colocasse em pauta segurança, transparência, legalidade e ética de algumas dessas soluções desenvolvidas até o momento. Com esse pretexto nasceu a General Data Protection Regulation (GDPR) cujo objetivo era englobar todos esses princípios de proteção e privacidade e posteriormente surgindo dessa base a Lei Geral de Proteção de Dados (LGPD), uma regulação adaptada para a realidade brasileira. Uma corrida se iniciou por parte das empresas para se adequarem, pois, o não cumprimento dessas leis geram aplicações de penalidades com valores elevados. Esse ocorrido virou objeto de estudo, e como resultado mostrou que por conta dessa corrida, houve uma série de problemas nas implantações desses princípios por parte dessas empresas sendo elas de pequena ou larga escala. Este trabalho tem como objetivo disponibilizar palavras-chave no contexto da GDPR, LGPD, e ética IA que auxiliem desenvolvedores na implementação dessas regulações, usando como base *issues* de empresas de código aberto. Como resultado implementar uma tabela com esses principais conceitos abordados na prática, auxiliando futuros desenvolvedores na realização dessa tarefa. Para isso usamos uma metodologia utilizando técnicas de Mineração de Dados (MD) sobre esses projetos disponibilizados em plataformas de versionamento para que as *issues* fossem extraídas e as palavras-chave fossem classificadas ao final de uma validação de precisão. Como conclusão deste trabalho é esperado que os resultados obtidos sejam de grande utilidade para os desenvolvedores no auxílio da implementação dessas questões, além de mostrar uma certa tendência e comportamento por parte dos projetos nessas fases de implantação.

Palavras-chave: LGPD, GDPR, Ética, Código Aberto, Mineração de Dados, Inteligência Artificial

Abstract

With the development of technology, user data has become one of the most important assets for companies, followed by the emergence and uncontrolled use of Inteligência Artificial (IA) put the privacy of these users at risk. Given this scenario, the need arose to create a regulation that put safety, transparency, legality and ethics on the agenda of some of these solutions developed so far. With this pretext, General Data Protection Regulation (GDPR) was born, whose objective was to encompass all these principles of protection and privacy and later emerging from this base the General Data Protection Law (Lei Geral de Proteção de Dados), a regulation adapted to the Brazilian reality. A race has started on the part of companies to adapt, since non-compliance with these laws generates applications of penalties with high values. This event became an object of study, and as a result it showed that because of this race, there were a series of problems in the implementation of these principles by these companies, whether small or large scale. This work aims to provide keywords in the context of GDPR, LGPD, and Ethics IA that help developers in these implementations, using *issues* from open source companies as a basis. As a result, implement a table with these main concepts covered in practice, helping future developers to carry out this task. For this, we used a methodology using Mineração de Dados (MD) techniques on these projects available on versioning platforms so that the *issues* were extracted, and the keywords were classified at the end of a precision validation. As a conclusion of this work, it is expected that the results obtained will be of great use to developers in helping the implementation of these issues, in addition to showing a certain tendency and behavior on the part of the projects in these phases of implementation.

Keywords: LGPD, GDPR, Ethical, Open Source, Data Mining, Artificial Intelligence

Sumário

1 Introdução **1**
1.1 Problema de Pesquisa 2
1.2 Justificativa 3
1.3 Objetivo Geral 4
1.4 Resultados Esperados 4
1.5 Metodologia de Pesquisa 4
1.6 Estrutura da Monografia 5

2 Embasamento Teórico **7**
2.1 General Data Protection Regulation (GDPR) 7
2.1.1 Princípios 7
2.2 Lei Geral de Proteção de Dados (LGPD) 8
2.2.1 Fundamentos 9
2.2.2 Princípios 10
2.3 Inteligência Artificial 11
2.3.1 Diretrizes Éticas em Inteligência Artificial 12
2.4 Data Mining 13
2.4.1 Técnicas e Algoritmos 13
2.5 Processamento de Linguagem Natural 14
2.6 Trabalhos Relacionados 15

3 Metodologia **17**
3.1 Base de Dados 18
3.2 Extração de Dados 19
3.3 Tratamento de Dados 22
3.4 Classificação de Dados 22
3.5 Geração da Tabela 27

4 Resultados **30**
4.1 Análise de Erros, Tendências e Comportamento 30

4.2 Verificação dos Resultados da Tabela com o Contexto das Regulações . . . 33
4.3 Ameaças e Limitações para Verificação 34

5 Conclusão **36**

Referências **38**

Lista de Figuras

1.1 Fases da Extração, Tratamento, Classificação e Geração da Tabela. 5

3.1 Diagrama completo da metodologia usada para criação da tabela automatizada. 17
3.2 Organização do DataBase. 18
3.3 Processo de Mineração do CodeTriage®. 20
3.4 Processo de Mineração do GitHub®. 21
3.5 Processo de Tratamento de Dados. 22
3.6 Palavras Escolhidas nos princípios da GDPR[1], LGPD [2] e Diretrizes Éticas IA [3] para Classificação de Dados. 23
3.7 Processo de Classificação de Dados. 24
3.8 Gráfico de Sinônimos do Princípio (*Security*) da GDPR de Issues Open. . 25
3.9 Gráfico de Sinônimos do Princípio (*Security*) da GDPR de Issues Closed. 26
3.10 Princípios GDPR, LGPD e Diretrizes Éticas IA da Issues Open dos Projetos. 26
3.11 Princípios GDPR, LGPD e Diretrizes Éticas IA da Issues Closed dos Projetos. 27

4.1 Resultado de Classificação da GDPR com Similaridade 75%. 31
4.2 Resultado de Classificação da IA com Similaridade 95%. 31

Lista de Tabelas

3.1 Resultado do Site com DataSets. 19
3.2 Resultado dos Tópicos com DataSets no GitHub®. 19
3.3 Resultado dos Processos de Extração do CodeTriage® e GitHub®. 21
3.4 Resultado GDPR. 28
3.5 Resultado LGPD. 29
3.6 Resultado IA. 29

4.1 Quantidade de Citações por Regulação. 32
4.2 Quantidade de Citações por Categoria de Issue. 32
4.3 Verificação Palavras-chave vs Issues. 34

Lista de Abreviaturas e Siglas

AM Aprendizado de Máquina.

ANPD Autoridade Nacional de Proteção de Dados.

GDPR General Data Protection Regulation.

IA Inteligência Artificial.

KDD Descoberta de conhecimento em bancos de dados.

LGPD Lei Geral de Proteção de Dados.

MD Mineração de Dados.

NLTK Natural Language Toolkit.

PLN Processamento de Linguagem Natural.

Capítulo 1

Introdução

Um projeto pode enfrentar muitos problemas derivados da ausência de elementos como segurança e ética em seu desenvolvimento, levando a instabilidade da aplicação, atrasos na entrega, complexidade na usabilidade e falta de confiabilidade por parte dos usuários. Tornou-se assim um objeto de estudo a ausência desses elementos tanto no meio acadêmico quanto em empresas de modo a buscar melhores soluções para esses desafios [4].

A preocupação com a ética à medida que sistemas Inteligência Artificial (IA) tomam cada vez mais decisões em setores da vida humana vem trazendo preocupações e danos à sociedade junto de benefícios econômicos, exemplo disso temos: entrevistas de emprego decididas por sistemas automatizados [5] como preocupações e design de produtos sendo desenvolvidos por IA [6] como benefícios. Temos também reportagens que mostraram empresas privadas recorrendo a softwares IA sem supervisão do Governo para decidir questões sobre saúde e medicina, emprego, financeiro e judiciário sem garantias de que aquele programa possua imparcialidade [7]. A partir disso a ética em IA está preocupada em determinar qual ação ou política seria melhor, dado um conjunto particular de circunstâncias [8].

Diante disso, equipes de desenvolvimento de software partiram em busca de novos métodos [9, 10], metodologias [11, 12] e frameworks [13, 14] para apoiar elicitação dos requisitos éticos legais. Com o surgimento do conceito de Mineração de Dados (MD)[15], Processamento de Linguagem Natural (PLN) [16, 17] atrelado ao Aprendizado de Máquina (AM)[18, 19], as equipes de desenvolvimento identificaram um grande potencial no uso dessas tecnologias para aprofundar os estudos de tal forma a conseguir requisitos que, em simultâneo, fossem funcionais e eticamente corretos [15].

Desde 25 de maio de 2018 a General Data Protection Regulation (GDPR) criada com objetivo de regulamentar o controle e proteção da privacidade dos dados, vem aplicando multas pesadas em empresas que infringissem os princípios estipulados na proteção de

dados. Por ser uma das companhias líder em multa recorde [20], a Google buscou rapidamente mudanças em suas soluções.

A Lei Geral de Proteção de Dados (LGPD) apesar de ter entrado em vigor em 2020, dois anos após a GDPR, se preocupou com o tratamento e armazenamento das informações das pessoas com a ética na integridade desses dados[21]. Canedo et al. [22] propuseram um modelo usando Business Process Modeling Notation (BPMN) para apoiar as organizações brasileiras na implementação da LGPD. Martins et al. [23] investigou e modelou uma análise Formal Concept Analysis (FCA) da regulação brasileira, implicando em uma racionalização sobre os conceitos, atributos e implicações ocultas na LGPD.

Encontramos ética na maioria dos fundamentos da LGPD, o respeito à privacidade (Art. 2.º I), a inviolabilidade da intimidade, da honra e da imagem (Art. 2.º IV) e a preocupação com os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais (Art. 2.º VII) [2].

Nesse trabalho com essas tecnologias em mente e com auxílio de trabalhos já realizados nessa área [17], [24], [25], sera feito um levantamento usando MD de termos citados no contexto dos princípios da GDPR e LGPD e as diretrizes éticas em IA que são mais utilizadas nos repositórios de código aberto do mercado e construir uma tabela automatizada contendo palavras-chave no contexto dessas regulamentações para auxiliar os desenvolvedores na implementação de futuras soluções com esses princípios e diretrizes.

1.1 Problema de Pesquisa

Em um sistema de IA com funcionamento de tentativa e erro, a interpretação dos dados coletados de um indivíduos é usada para decidir a melhor ação a ser tomada, a ética se torna um problema, pois se torna o centro das atenções dessas empresas como Alibaba [26], Netflix [27], Google [28]. A IA tem como um dos objetivos, regular e decidir a vida das pessoas usada em benefício da sociedade, além de proteger os direitos humanos e respeitar a privacidade e autonomia desses indivíduos [29].

Um problema que vem chamando a atenção dos usuários é sobre alguns questionamentos em relação ao comportamento de algumas organizações. Baker-Brunnbauer et al. [30] apresentou os seguintes questionamentos: *Que categoria de consciência tem o gerenciamento destas empresas de IA ao abordar o impacto social do produto ou serviço que oferecem? Quais diretrizes e princípios éticos conhecidos foram usados no desenvolvimento?*, o viés humano dessas questões abordadas são consideradas fatores importantes no treinamento de um algoritmo IA.

A reputação temível da GDPR de que *"a lei que pode multar você em €20 milhões"* vai levando designers ao entendimento errado do que realmente seria seus princípios e gerando

problemas a respeito de certos mitos como: *"Interrompe o processamento de dados"*, *"Fala apenas sobre o usuário"* e por fim *"É tudo questão de consentimento"*[31].

A implementação da recente LGPD até então baseada na GDPR, pode impactar as equipes de desenvolvimento, pois é necessário um nível de conhecimento e uma certa experiência por parte dos desenvolvedores sobre a regulação. A implementação em conjunto com metodologias ágeis enfrentam problemas com requisitos de software por falta de conhecimento por parte dos profissionais [32].

Um estudo apresentou que muitos desenvolvedores não se sentiam capazes de desenvolver sistemas que seguiam os princípios de privacidade, e por conta dessa falta de familiaridade, obstáculos foram gerados na integração da solução com a GDPR [33] e essa falta de conhecimento na integração criou problemas éticos na engenharia de requisitos [34] do projeto.

Após esse ocorrido os profissionais de Tecnologia da Informação e Comunicação (TIC) que tinham certo estudo, afirmaram que o conhecimento deles em relação a GDPR não era o suficiente por indisponibilidade de ferramentas online e indisponibilidade de suporte por parte de instituições aos seus colaboradores [33].

1.2 Justificativa

A justificativa desse trabalho tem como base os acontecimentos [31, 32, 29] envolvendo essas regulações visando evitar que pequenas e grandes empresas sofram com penalidades aplicadas com valores astronômicos[20], alem de auxiliar esses profissionais no projeto.

Para que desenvolvedores experientes e inexperientes consigam superar os obstáculos gerados na implementação e assim terem melhor aproveitamento com as diretrizes da GDPR[33] e LGPD[23].

Para que desenvolvedores tenham conhecimento sobre a política de privacidade dos dados[2] e consigam conciliar com metodologias ágeis, com intuito de pôr fim aos mitos criados a respeito dos princípios, utilizando tecnologias acessíveis [31], bem como proteger os direitos humanos e respeitar a privacidade e autonomia desses indivíduos em sistemas em IA[29].

Nesse contexto este trabalho está diretamente ligado ao uso da tecnologia no uso da MD desses princípios e diretrizes, para auxiliar os profissionais de TIC na superação desses obstáculos [33], através da disponibilização de uma tabela contendo palavras-chave a respeito desses contextos, para proteger os direitos e respeitar a privacidade das informações dos usuários[29].

1.3 Objetivo Geral

O objetivo deste trabalho é disponibilizar palavras-chave citadas em repositórios de software de código aberto no contexto de princípios da GDPR, LGPD, e ética IA para auxiliar desenvolvedores na implementação dessas políticas. Para atingirmos o objetivo deste trabalho, os seguintes objetivos específicos foram definidos:

- Uso do MD para mineração dos repositórios de software e dados encontrados das *issues*.
- Uso de algoritmos de PLN para identificar palavras ligadas aos respectivos princípios e diretrizes éticas.
- Elaborar uma tabela automatizada com essas palavras-chave no contexto das regulações para ser usadas como ferramenta de auxílio pelos desenvolvedores e outros profissionais.
- Verificar o relacionamento do contexto entre as palavras-chave e *issues* dos projetos analisados, para saber se estão alinhados.

1.4 Resultados Esperados

O resultado desse trabalho será a divulgação de uma tabela de palavras-chave gerada de forma automatizada, contendo termos no contexto dos princípios e diretrizes éticas para ser usada por desenvolvedores e profissionais na implementação dessas regulações.

1.5 Metodologia de Pesquisa

Usando como base algumas abordagens feitas em estudos sobre engenharia de requisitos com AM, onde as fases eram compostas por: *Normalização, Extração, Seleção, Classificação e Desempenho* [35]. Ou separadas em conjuntos de objetivos como: *Aplicação de Técnicas de Mineração de Texto e Desenvolvimento de Modelos Preditivos* [36].

A Figura 1.1 apresenta a estratégia de pesquisa adotada neste trabalho. Nesse trabalho foi reunido os pontos mais relevantes desses métodos para a concepção e criação de uma ferramenta para minerar essas palavras-chave no contexto desses princípios e diretrizes tendo como estipulado as seguintes fases:

1. **Base de Dados:** Etapa onde é feita a seleção e escolha da plataforma que será minerada, além de definir como levantar esses projetos de software de código aberto.

2. **Extração de Dados:** *Crawlers* vão ser responsáveis por pegar e varrer os links desses repositórios em questão para obtenção dos dados textuais contidos nas *issues* de cada um.

3. **Tratamento de Dados:** Com os dados brutos obtidos anteriormente faremos a *Tokenization*, removendo símbolos, pontuações e *Stop-Words*.

4. **Classificação de Dados:** Fase de aplicação dos filtros e técnicas de MD para separar as palavras-chave em seus devidos princípios GDPR, LGPD e diretrizes éticas IA.

5. **Geração da Tabela:** Fase do processo onde geramos uma tabela automatizada com todas as palavras-chave separadas por princípios e diretrizes.

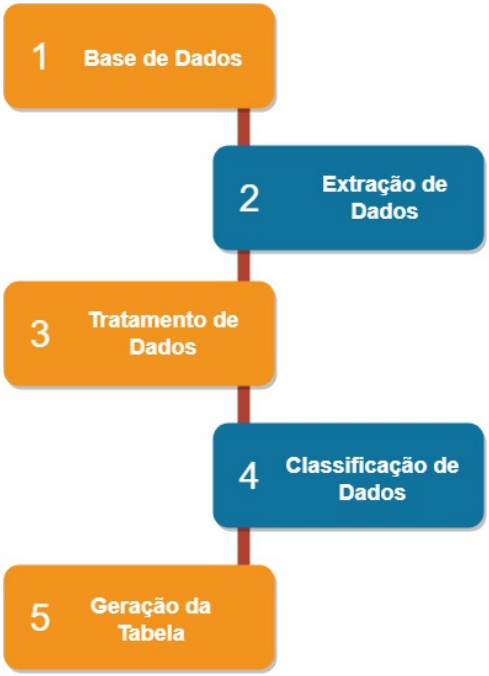

Figura 1.1: Fases da Extração, Tratamento, Classificação e Geração da Tabela.

1.6 Estrutura da Monografia

Esta monografia será estruturada da seguinte maneira: No Capítulo 2 é realizado o embasamento teórico sobre os principais conceitos da GDPR, LGPD e Diretrizes Éticas em IA no design, PLN, e MD.

No Capítulo 3 sera apresentadas as buscas por plataformas e alternativas de levantamentos do projeto de código aberto, além de apresentar a construção, fluxograma do planejamento dos módulos e do funcionamento da construção da tabela.

No Capítulo 4 serão apresentados os resultados da formação da tabela automatizada, uma discussão sobre a análise desses resultados junto as ameaças para validação do trabalho.

No Capítulo 5 serão apresentados a conclusão com a contribuição desse trabalho seguido de uma discussão sobre trabalhos futuros relacionados com esse tema.

Capítulo 2

Embasamento Teórico

Neste capítulo serão abordados os principais termos, conceitos e tecnologias relevantes para a discussão e construção da tabela de palavras-chave ao final do trabalho.

2.1 General Data Protection Regulation (GDPR)

A regulação europeia (EU) 2016/679, GDPR foi sancionada no dia 27 de abril de 2016 e começou a ser aplicada em toda Europa em 25 de maio de 2018[1].

A GDPR foi um marco dos últimos anos na União Europeia, pois substituiu a Diretiva de Proteção de Dados de 1995, que tinha como objetivo a proteção de indivíduos no que diz respeito ao processamento de dados pessoais e a livre circulação deles. Entre 2011 e 2012 surgiu preocupações e a necessidade do fortalecimento da proteção desses dados na comunicação, principalmente no fortalecimento dos direitos de privacidade online para impulsionar a economia digital na Europa. Em 2014 o Parlamento europeu decidiu com unanimidade forte apoio em adotar o projeto da GDPR.

No ano seguinte em 2015 o conselho europeu chegou a uma abordagem geral sobre a GDPR, seguido de recomendações por parte da Autoridade Europeia para a Proteção de Dados (AEPD), para assim gerar o plano de ação da implementação da GDPR no início de 2016, e com isso ter sua publicação no dia 27 de abril de 2016 e por fim entrar em vigor em 2018 [37].

2.1.1 Princípios

A partir de alguns conceitos e acordos definidos por parte dos países membros foram definidos princípios para o tratamento e circulação desses dados pessoais sendo eles:

1. Legalidade, Justiça e Transparência: os dados devem ser processados de forma lícita, justa e transparente em relação ao titular.

2. Finalidade: os dados devem possuir propósitos legítimos, específicos, explícitos e informados ao titular.

3. Minimização de Dados: os tratamentos devem ser adequados, relevantes e limitados ao necessário em relação às finalidades para as quais são processados.

4. Exatidão: os dados devem ser precisos e se necessário, atualizados, caso não sejam devem ser tomadas todas as medidas razoáveis para assegurar que os dados pessoais inexatos, sejam apagados ou retificados sem demora.

5. Armazenamento: os dados devem ser mantidos de uma forma que permita a identificação dos titulares por não mais do que o necessário para as finalidades aos quais são processados. Os dados pessoais podem ser armazenados por períodos mais longos, exclusivamente para fins de arquivamento de interesse público, fins de pesquisa científica, histórica ou estatísticos, conforme o (*art. 89.º n.º 1*).

6. Integridade e Confidencialidade: os dados são processados para garantir a segurança adequada dos dados pessoais, incluindo proteção contra processamento não autorizado ou ilegal e contra perda, destruição ou dano acidental, usando medidas técnicas ou organizacionais apropriadas.

O responsável pelo tratamento dos dados deve cumprir os princípios listados acima sobre pena e sanções previstas no (*art. 83.º*) caso não seja feito [1].

2.2 Lei Geral de Proteção de Dados (LGPD)

A regulação brasileira 13.709/2018, LGPD foi sancionada em 14 de agosto de 2018, mas passou a valer em 18 de setembro de 2020, e as penalidades passaram a ser aplicadas a partir de 1 de agosto de 2021[2].

A proteção de dados foi consolidada em 1990 e estava diretamente ligado ao modelo de negócio da economia digital. Com o surgimento da GDPR [1] grandes bases de dados internacionais eram relacionadas às pessoas devido aos avanços tecnológicos e pela globalização.

A partir desse panorama internacional a LGPD foi um marco de grande impacto em instituições privadas e públicas no Brasil, por tratar da proteção dos dados desses indivíduos em qualquer categoria de tratamento de informações que utilizassem dados pessoais, seja pessoa natural ou jurídica [38].

Apesar da LGPD ser considerada moderna, uma abordagem racional acerca do tema de privacidade de dados já vinha sendo abordada por outras regulações e normas. A (ISO 27 001) possui certa semelhança ao que temos hoje na LGPD, devido a diversas discussões

com afirmativas falsas foram criadas como pretexto de *"Aqueles com certificação ISO 27 001 estão automaticamente adequados à LGPD"*(2021, p. 114) [39]. Essa afirmação é falaciosa, pois a (ISO 27 001) era uma norma que cuidava da segurança da informação num âmbito mais geral da (TI).

Cuidar e proteger essa informação é uma das prioridades para o setor de TI com alguns conceitos e boas práticas de segurança e governança dos dados, com isso organizações fora do setor de TI de certa forma começam a fazer uso dessas ferramentas e assim seguirem sua jornada de conformidade com as leis e regulações dos dados com a LGPD [39].

Adequações feitas por essas organizações com a LGPD avançou consideravelmente, muitos projetos de adequação foram finalizados em tempo recorde com qualidade baixa e outros implementaram medidas paliativas insuficientes para evitar a multa.

As sanções administrativas previstas pela LGPD e fiscalizadas pela Autoridade Nacional de Proteção de Dados (ANPD) podem resultar na quebra de um negócio dependendo da gravidade. No entanto, evitar essas sanções legais não deveria ser o foco desses projetos, pois seria como utilizar um cinto de segurança por medo de levar uma multa e não por ser um item de segurança que pode salvar sua vida [40].

Diversos desafios surgem à organização ao se adequar à LGPD, iniciando a jornada a empresa deve compreender o conceito de dados pessoais sensíveis e não sensíveis.

Essa proposta de conceito gera diversas confusões jurídicas por não seguir uma lógica além de tratar diversas características de um indivíduo, sendo um desafio de alinhamento com questões éticas para a proteção desses dados pessoais, sendo necessário cumprir alguns requisitos como a prevenção de danos, o enfrentamento da desigualdade informacional, as injustiças, discriminações e a autonomia dos indivíduos são apenas algumas dos requisitos necessários para o tratamento das questões éticas para superar estes desafios [41].

2.2.1 Fundamentos

Os fundamentos da LGPD abrangem tanto os meios físicos quanto digitais, as pessoas físicas e jurídicas com direito público ou privado, além de proteger os direitos fundamentais de liberdade, privacidade e livre personalidade da pessoa física, para isso ser possível a LGPD tem como fundamentos [2]:

I — Respeito à privacidade.

II — Autodeterminação informativa.

III — Liberdade de expressão, de informação, de comunicação e de opinião.

IV — Inviolabilidade da intimidade, da honra e da imagem.

V — Desenvolvimento econômico, tecnológico e a inovação.

VI — Livre iniciativa, a livre concorrência e a defesa do consumidor.

VII — Direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais.

A garantia do respeito a privacidade (I) foi criada para que todos os dados não públicos de um indivíduo fossem invioláveis, a autodeterminação Informativa (II) do que deve ou não ser feito com os dados, o conhecimento em saber quais e como essas informações vão ser utilizadas por empresas devem ser repassadas aos indivíduos. Os fundamentos de desenvolvimento econômico, tecnológico, inovação(V) e a livre iniciativa, concorrência e defesa do consumidor(VI) são direcionados ao desenvolvimento e interesse nacional [42].

A LGPD possui fundamentos repetidos previstos na Constituição Federal para garantir e reforçar sua aplicabilidade. Cada fundamento possui devida importância deixando evidente a preocupação em proteger os direitos fundamentais dos indivíduos, resultando em fundamentos relacionados à alguns artigos da constituição do Brasil (*art. 3°, 4°, 5°, 7°,219°*)[43], [38].

2.2.2 Princípios

A boa-fé no tratamento de dados pessoais é o primeiro critério a ser cumprido por parte das organizações na implementação da LGPD. A partir disso uma série de princípios foram queridos garantir uma efetividade no tratamento desses dados, sendo eles [44]:

I — Finalidade: os dados devem possuir propósitos legítimos, específicos, explícitos e informados ao titular.

II — Adequação: o tratamento deve ter compatibilidade com as finalidades informadas ao titular, conforme o contexto do tratamento.

III — Necessidade: limitação do tratamento ao mínimo necessário para a realização de suas finalidades.

IV — Livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais.

V — Qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados.

VI — Transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial.

VII — Segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão.

VIII — Prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais.

IX — Não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos.

X — Responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

Esses princípios são importantes para o tratamento dos dados por parte das instituições, se a organização segue todos os princípios da LGPD as chances de uma sansão administrativa aplicada por parte da Autoridade Nacional de Proteção de Dados (ANPD) são minímas [40].

2.3 Inteligência Artificial

A IA ao longo da história mostrou ser uma ferramenta útil para a computação. Os primeiros estudos sobre essa tecnologia surgiram na década de 1940 com foco na Segunda Guerra Mundial. Por possuir um grande poder de processamento acabou sendo usada em análise balística, quebra de criptografias e cálculos complexos. Após o fim da guerra o computador deixou de ser de uso restrito militar e passou a ser utilizado por empresas e universidades, tendo assim a disseminação desta tecnologia [45].

De forma geral a IA é definida segundo os livros como *"Inteligência Artificial envolve utilizar métodos baseados no comportamento inteligente de humanos e outros animais para solucionar problemas complexos."* Coppin et al. [46] mostraram que, com apenas um computador com capacidade suficiente de processamento, fornecendo o mínimo de inteligência é possível criar uma máquina que seja inteligente e consciente assim como um ser humano.

Diversas estratégias foram usadas pelos pesquisadores para tentar identificar o significado da IA, com isso houve uma categorização de 4 principais pensamentos em volta deste tema ao longo dos anos sendo: *pensamento como um humano, pensamento racional, agindo como seres humanos, agindo racionalmente.* Essas categorias de abordagem envolvem uma ciência empírica com hipóteses e confirmação experimental [47].

Ao realizar estudos sobre limites que podem ser alcançados usando IA, analises sobre o quanto de similaridade com seres humanos é possível chegar e outras indagações foram

se tornando comum. Para que esses estudos sejam feitos um grande fluxo de dados são necessários para processamento, a partir disso levantou um alerta sobre até que ponto essas pesquisas podem ser consideradas moralmente corretas [48].

A IA com suas técnicas vem sendo um grande aliado da engenharia de requisitos pois vem auxiliando atividades que demandam tempo e possuem abundância de recursos Aguilar et al. [49] realizaram um mapeamento sistemático e mostraram diferentes cenários de uso da IA em vários contextos, além disso, os autores divulgaram as recentes áreas de atuação da IA, seus resultados obtidos e possíveis trabalhos que podem surgir.

Investimentos em IA foram feitos em várias áreas como agricultura [50, 51], saúde [14], análise de risco financeiro [52, 53] além de áreas da (TI) como engenharia de requisitos [34, 18, 24]. Demostrando como resultado um avanço global por parte do uso dessa tecnologia. Mas estudos mostram alguns riscos sistêmicos que ela traz consigo como falhas em cascatas em alguns algoritmos, fazendo com que empresas tomem certo cuidado ao fazer uso [47].

2.3.1 Diretrizes Éticas em Inteligência Artificial

O uso da IA para criar máquinas e algoritmos pensantes vem sendo objeto de estudo em diversas obras literárias, com isso uma série de questões éticas vem sendo abordadas para garantir que não prejudiquem humanos. O objetivo é que sua implantação na sociedade seja moralmente ética, um algoritmo de AM que realiza avaliações de perfil econômico do indivíduo antes de conceder um empréstimo, é apenas um de muitos exemplos considerados desafios éticos no cotidiano [54].

A ética em IA teve como base um estudo que vinha sendo proposto pela comunidade onde a ética alinhada ao design surgiu de um trabalho dedicado e consenso, visando articular princípios de alto nível para serem usados em sistemas autônomos e inteligentes [55].

O resultado desse trabalho e consenso dedicados foi a criação e uso de princípios gerais alinhados com o design com os (*Direitos Humanos*) internacionais para se ter respeito e proteção, (*Bem Estar*) humano como principal critério de sucesso no desenvolvimento, (*Agência de Dados*) com a capacitação de pessoas para compartilhamento de dados com segurança, (*Eficácia*) com as provas de adequação, (*Transparência*) em todos os processos, (*Responsabilidade*) fornecendo uma razão para as decisões tomadas, (*Conscientização*) sobre o uso indevido da IA e por fim a (*Competência*) onde é necessário conhecimento e habilidade para um desenvolvimento seguro e eficaz [55].

Partindo dos princípios básicos a respeito da ética alinhada ao design, um grupo chamado High-Level Expert Group on AI (AI HLEG) [3] escreveu um documento publicado

pela comissão europeia com diretrizes e princípios éticos com base nos direitos fundamentais.

Onde deve ser respeitado para garantir que os sistemas IA sejam implantados e usados de forma confiável, sendo eles o (*Respeito pela autonomia humana*) onde os humanos devem ser livres para escolher se querem interagir ou não com sistemas IA, (*Prevenção de danos*) ao qual o sistema IA deve prevenir e evitar que humanos tenham sua dignidade mental ou física protegida, (*Justiça*) no desenvolvimento de um sistema IA sendo feito de forma justa, (*Explicabilidade*) sendo crucial para que os usuários tenham confiança no sistema IA, através da transparência e da comunicação aberta de decisões.

2.4 Data Mining

Não podemos falar em MD sem citar os conceitos da Descoberta de conhecimento em bancos de dados (KDD).

O termo surgiu em 1989 por um grupo de pesquisados sendo definido em 1996 como: *"Um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dado"* [56].

Os autores definiram assim os seguintes componentes para a aplicação do KDD[56]:

1. o problema em que será aplicado o processo de KDD;
2. os recursos disponíveis para a solução do problema;
3. os resultados obtidos a partir da aplicação dos recursos disponíveis para a solução.

Os processos operacionais da KDD são divididos em (*Pré-processamento*) onde o objetivo é preparar os dados, (*Mineração de Dados*) onde é realizada a busca efetiva por informações e por fim o (*Pós-Processamento*) onde é feito a avaliação da utilidade do conhecimento descoberto. O MD é a parte principal do KDD, pois são definidos as técnicas e os algoritmos a serem utilizados no problema. A escolha da técnica está relacionada com a categoria de tarefa KDD que vai ser realizada [56].

2.4.1 Técnicas e Algoritmos

Existem diversas técnicas de MD como Rede Neural Artificial [57], Lógica Difusa [58], Algoritmo Genético [59], técnicas estatísticas entre outras. Logo, escolher a que melhor se adéqua ao contexto da sua implementação é crucial, pois essas técnicas são dividas em tipos sendo as: *Tradicionais, Especificas e Hibridas* [56].

As *Tradicionais* são muito utilizadas independente do contexto em que a mineração é feita por apresentarem bons resultados em aplicações KDD.

A *Rede Neural Artificial RNA* constrói um modelo automático simples baseado em um sistema neural biológico, através do aprendizado de padrões com repetidas apresentações dos dados na rede corrigindo o próprio erro [57].

A *Lógica Difusa* permite lidar com informações imprecisas ou subjetivas oferecendo flexibilidade na definição e avaliação de conceitos e por sua eficiência podendo ser incorporada em MD de banco de dados de larga escala como fundos de previdência sociais [58], recursos humanos de universidades e faculdades [60].

Baseada na seleção natural, o *Algoritmo Genético AG* trabalha com três componentes principais variáveis que descrevem o problema, as restrições indicando os valores de atribuições de variáveis e funções-objetivos responsáveis por avaliar o resultado. Essa técnica é indicada para problemas complexos como gerenciamento financeiro [59], aprimoramento de desempenho de rede a partir de dados de transmissão [61] essa abordagem pode não garantir a solução ótima, mas algo bem próximo ou aceitável para um problema.

As *Técnicas Estatísticas* possui diversas técnicas tradicionais muito comuns no meio estatístico que se mostraram muito eficientes para analisar e interpretar dados nas mais diversas áreas de estudo as mais conhecidas são *Análise de Regressão* [62], *Redes Baysianas* [63] e *Análise Exploratória de Dados* [64].

As *Especificas* são técnicas muito particulares desenvolvidas apenas para suprir tarefas de KDD. Um exemplo de uso específico de algoritmo é chamado (*Apriori*) onde é realizado são feitas buscas de conjuntos de itens frequentes em um conjunto de dados, sendo muito útil quando tentamos encontrar padrões periódicos em banco de dados de séries temporais [65].

A partir dele surgem variações como *DHP (Direct Hashing and Pruning)* usado para reduzir sucessivamente o tamanho da transação base de dados através de uma função de hash para geração do conjunto de itens candidatos, o *Partition* e *ParMaxEclat* [56].

No caso das *Híbridas*, as combinações tanto de *Tradicionais* quanto técnicas *Especificas* podem formar e definir sistemas híbridos de MD, trazendo vantagens como sinergia entre os algoritmos combinados gerando assim um poderoso sistema e com menos falha, possibilitando construir sistemas (*Híbridos Sequenciais*) onde uma técnica atua sobre a saída da outra e (*Híbrido Auxiliar*) a técnica A utiliza a B para gerar seus resultados [66].

2.5 Processamento de Linguagem Natural

A PLN é uma poderosa ferramenta quando tratamos problemas que envolvem dados de escrita humana.

A análise sintática no campo da PLN sempre foi um tópico importante para pesquisadores poderem criar tarefas como perguntas e respostas, compreensão de *strings* na

pesquisa, fazer uma análise semântica para construir uma base de conhecimento e a partir disso usar com outras técnicas de MD como redes neurais [67].

Uma das vantagens da aplicação da PLN é a capacidade de processamento envolvendo dados de texto não padronizados, se tornando um grande aliado em estudos com foco na área da saúde através do processamento de dados clínicos armazenados em documentos não padronizados em centros clínicos contendo informações cruciais para estudo sobre medicamentos, diagnósticos e possíveis complicações de pacientes, tendo sempre como preocupação a privacidade e segurança desses indivíduos para estudo [68].

Uma de muitas desvantagens da PLN são possíveis falhas de interpretação decorrente da falta de robustez no sistema na fase inicial de implementação. Mesmo com essa falha não impediu o avanço da comunidade em grandes melhorias nas tarefas de PLN, como assistentes virtuais, reconhecimento de fala e analises de sentimento apresentado por um estudo que buscou de forma estruturada e resumida em várias dimensões a respeito de robustez na PLN [69].

Trazendo a PLN para o ambiente de TI, temos uma ampla aceitação de artefatos extraídos de *users stories* para o levantamento de requisitos do usuário no desenvolvimento ágil, e com isso um levantamento realizado por um estudo nas principais plataformas de divulgação acadêmica *SCOPUS, ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink e Google Scholar* mostrou que de 718 artigos publicados entre 2009 e 2020 com tema de elicitação de requisitos 38 deles tinham como abordagem o PLN [70].

2.6 Trabalhos Relacionados

Pesquisa em repositórios de software de mineração é uma pesquisa envolvendo seres humanos, pois os repositórios contêm geralmente dados sobre as interações dos desenvolvedores e usuários entre si. As questões éticas levantadas por tais pesquisas, apresentaram quatro estudos de caso que ilustram questões éticas típicas encontradas em projetos e como as considerações podem moldar os projetos antes que eles comecem [25].

A usabilidade de um usuário gera uma trilha de atividades durante o uso de aplicativos e softwares, logo proteger esses dados vem se tornando cada vez mais desafiador com a necessidade da preservação da política de privacidade dessas pessoas, os regulamentos existentes fornecem apenas uma abordagem de alto nível dificultando a interpretação por parte dos engenheiros de software, então uma pesquisa que forneça de forma estruturada esses requisitos de privacidade, usando a GDPR e o ISO/IEC 29 100 e de análise usando AM para extrair dados de código aberto (Google e Moodle) é sempre bem-visto [15].

A pesquisa no estágio inicial de uma aplicação requer um estudo profundo e rigoroso de grandes volumes de documentação, com os avanços do AM em PLN podem auxiliar para

extrair requisitos funcionais de documentações extensas. Mostrando que redes neurais, podem ser utilizados como uma ferramenta que automatiza essa fase inicial facilitando o trabalho dos stakeholders [19].

As histórias de usuários são aceitas como artefatos para capturar os requisitos do usuário no desenvolvimento ágil de software, pois são textos curtos em formato semiestruturado que expressam alguns requisitos. As técnicas de processamento de linguagem natural (PLN) oferecem uma vantagem potencial em aplicativos de história de usuário por realizarem uma revisão sistemática da literatura para capturar o estado atual da pesquisa em PLN sobre histórias de usuários [70].

Na última década, abundantes dados clínicos foram disponibilizados em formato computável em sistemas de Registro Eletrônico de Saúde (RES) e mídias sociais. Muitos sistemas de (RES) usam estruturas de documentos não padronizadas para registrar informações críticas sobre medicamentos, diagnósticos e possíveis complicações, além disso, dados de mídia social como tweets geralmente sofrem de alta variação de idioma. Os métodos de ciência de dados e processamento de linguagem natural (PLN) recentemente populares, podem desbloquear informações do texto narrativo e receberam grande atenção no domínio médico [68].

Capítulo 3

Metodologia

Neste capítulo apresentamos e utilizamos a metodologia proposta na Seção 1.5, para a geração e verificação da nossa tabela sobre LGPD/GDPR e Diretrizes Éticas em IA. A Figura 3.1 apresenta todos os passos e técnicas aplicados em cada parte da Metodologia.

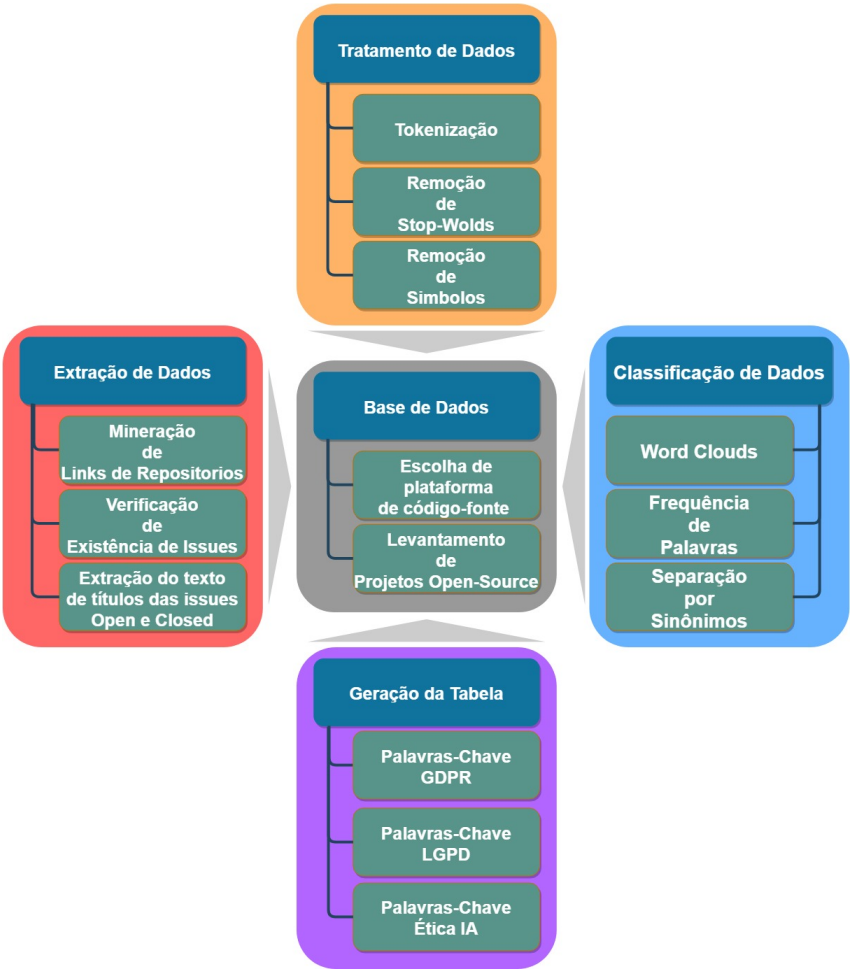

Figura 3.1: Diagrama completo da metodologia usada para criação da tabela automatizada.

Conforme a Figura 3.1 acima, o objetivo é ilustrar o funcionamento da metodologia que será implementada nos tópicos seguintes. A base de dados se encontra no centro com as demais fases direcionadas para ela, porque serve para demostrar que em todos os processos dentro de cada fase a base de dados é consultada.

3.1 Base de Dados

Sendo o ponto de partida de nossa metodologia, a base de dados usada neste trabalho é a parte mais importante de toda estrutura do projeto porque no *DataBase* conforme a Figura 3.2 armazenamos os dados do MD sendo: *URLs, Issues Open e Closed*. Dentro do nosso *DB Extract*, os resultados do PLN tanto na parte de Tratamento de dados com *Issues Open e Closed Clear* em *DB Treatment* e *WordClouds, Word Frequency* no *DB Classification* e por fim nossa Tabela gerada *DB Tabela* com os princípios da *GDPR, LGPD* e *Diretrizes Éticas IA*.

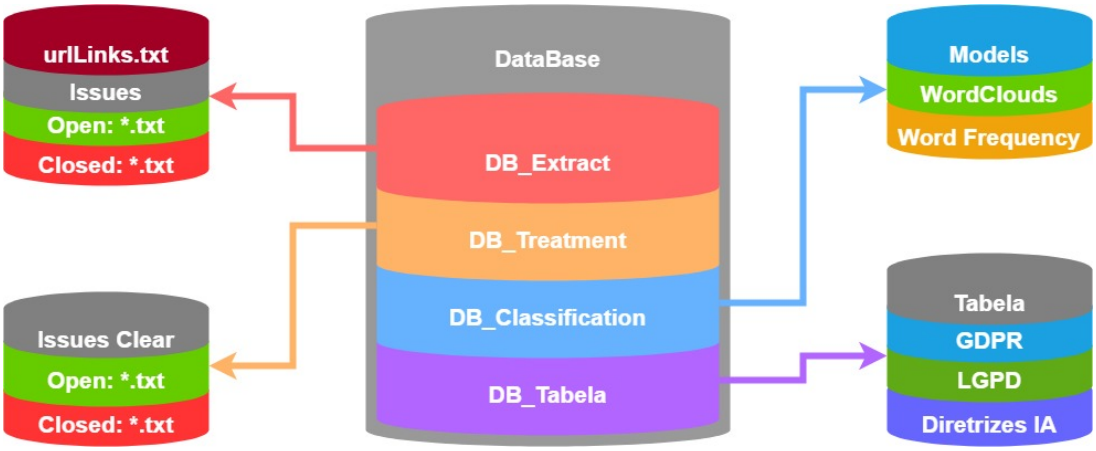

Figura 3.2: Organização do DataBase.

Partindo para o levantamento do *dataset* que será utilizado estabelecemos certas características para que os resultados da tabela fossem coletados. Ser populoso e diversificado na variedade de projetos, foram as principais características escolhidas para garantir a qualidade ao fim do trabalho.

A escolha e uso do GitHub® [71] como plataforma de hospedagem de código-fonte e arquivos com controle de versão é usada neste trabalho como levantamento de *issues*. Para que nossas características de *população e diversidade* fossem satisfeitas, foi feito um levantamento no Google® de *datasets* disponíveis para obter esses projetos de código aberto. Como resultado encontramos sites conforme a Tabela 3.1, além de tópicos Tabela 3.2 no próprio GitHub®.

Seguindo as características estipuladas na escolha do *dataset*, serviu para mostrar que a *Diversidade* foi o critério mais cumprido tanto pelos sites conforme a Tabela 3.1 quanto pelos tópicos de pesquisa no GitHub® conforme a Tabela 3.2.

Um diferencial proposto para o trabalho, foi a não obrigatoriedade de login para obtenção desses dados, pois em caso do ambiente necessitar desse procedimento pode impactar o MD.

Tabela 3.1: Resultado do Site com DataSets.

Sites	Diversidade	Login	Total de Projetos
24pullrequests[72]	Sim	Sim	508
up-for-grabs[73]	Sim	Sim	986
codetriage[74]	Sim	Não	5250
firstcontributions[75]	Sim	Não	Não Informado

Tabela 3.2: Resultado dos Tópicos com DataSets no GitHub®.

Tópicos	Diversidade	Login	Total de Projetos
beginner-friendly[76]	Sim	Não	1546
easy[77]	Sim	Não	1073
first-timers-only[78]	Sim	Não	63

Com os resultados obtidos nas Tabelas 3.1 e 3.2 o melhor *dataset* para extrair os links dos projetos foi o CodeTriage® [74]. Este *dataset* foi selecionado por possuir uma quantidade grande de projetos e ser de mais fácil acesso por não ser necessário login e ainda possui a maior quantidade de projetos dos resultados observados. Este site é uma plataforma gratuita para o desenvolvedor procura soluções de código aberto para contribuir, este ambiente conta com diversos projetos sendo eles de grande e pequena escala, que ajudou a construir nossa base de dados de forma diversificada.

Com o *dataset* e fonte de levantamento dos repositórios definidos, seguindo as características estabelecidas, partimos para a extração dos dados desses ambientes, com recursos e ferramentas de MD.

3.2 Extração de Dados

Nessa etapa o objetivo foi a extração das informações contidas no CodeTriage® [74] e GitHub® [71] conforme apresentado nas Figuras 3.2 e 3.4. O primeiro passo da extração foi a escolha de qual tecnologia seria usada ao longo do trabalho para o MD em questão,

usamos a linguagem de programação Python[79] em conjunto com uma biblioteca chamada Beautiful Soup [80] específica para uso de Web Crawlers responsável por navegar e extrair essas informações.

O primeiro elemento a ser extraído são os links dos repositórios contidos no CodeTriage® [74], para isso executamos uma série de processos conforme apresentado na Figura 3.3.

A página inicial era composta por um conjunto de elementos *cards* paginados cujo conteúdo era título e descrição de algum repositório. Para obtermos os links desses elementos foi necessário entrar nesses *cards*, na página desses componentes continha algumas informações sobre o projeto assim como o link em questão, com o link extraído e armazenado em uma lista retornamos a página principal e repetimos os processos novamente até o limite de páginas disponíveis.

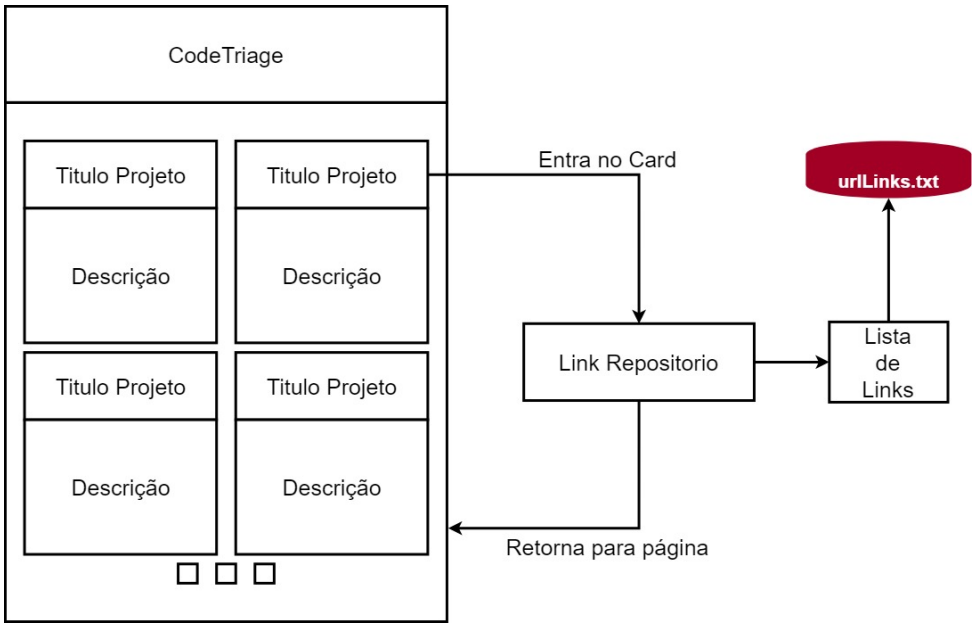

Figura 3.3: Processo de Mineração do CodeTriage®.

Ao final dos processos de MD e conforme apresentado na Figura 3.3, um número significativo de repositórios (Tabela 3.3) foi extraído. É importante lembrar que o site está em constante atualização, com isso nosso resultado pode variar em cada nova extração do CodeTriage® podendo ter adição ou remoção de alguns projetos.

Em relação aos processos envolvendo o GitHub®, conforme a Figura 3.4, foi usada a lista de links obtida anteriormente conforme a Figura 3.3 para percorrer a lista de repositórios seguindo uma série de condições.

O uso dessas condições tem o objetivo de garantir desempenho no processo do MD, entre elas estão a verificação da existência do projeto dentro do GitHub® e se este projeto de código aberto possui ou não *issues*. Cada *issue* sendo ela open ou closed tem o

conteúdo do título extraído e armazenado no respectivo arquivo (.Issue/Open/<Nome do Projeto>.txt) ou (.Issue/Closed/<Nome do Projeto>.txt).

Ao final do ciclo é levantado dados a respeito de quantos links foram realmente extraídos, quantos não possuíam *issue* ou não existiam, quantas *issues open* e *closed* foram extraídas no total.

Conforme apresentado na Tabela 3.3, ao fim das extrações de links no CodeTriage®, foram totalizados 5.484 links extraídos e passando para o GitHub® entre *issues open* e *closed* no total de 5.250. O valor se encontra abaixo do número obtidos, pois tivemos um total de 87 repositórios sem *issues*, além de 147 projetos 404 Not Found.

As causas para esse ultimo número pode ser por questões de disponibilidade por parte dos Owners ou de exclusão do projeto pelos mesmos.

Tabela 3.3: Resultado dos Processos de Extração do CodeTriage® e GitHub®.

	Total
Links Extraídos	5484
Repositórios Minerados	5250
Repositórios sem Issue	87
Repositórios não encontrados	147

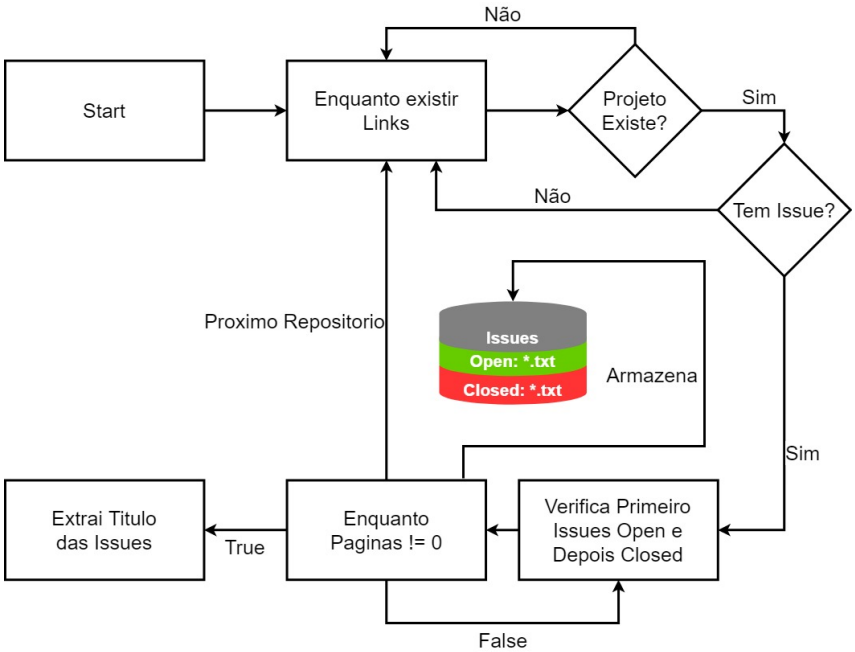

Figura 3.4: Processo de Mineração do GitHub®.

3.3 Tratamento de Dados

O tratamento desses dados é uma importante fase dentro do MD porque queremos maximizar o desempenho além de obter resultados mais concretos com a remoção de elementos textuais irrelevantes para a pesquisa, como a remoção de *stop words, pontuation e símbolos.* O processo de limpeza conforme descrito na Figura 3.5 percorre uma lista com os nomes de cada projeto, em seguida cada arquivo de *issue open* e *closed* do respectivo projeto é aberto para que a limpeza seja feita.

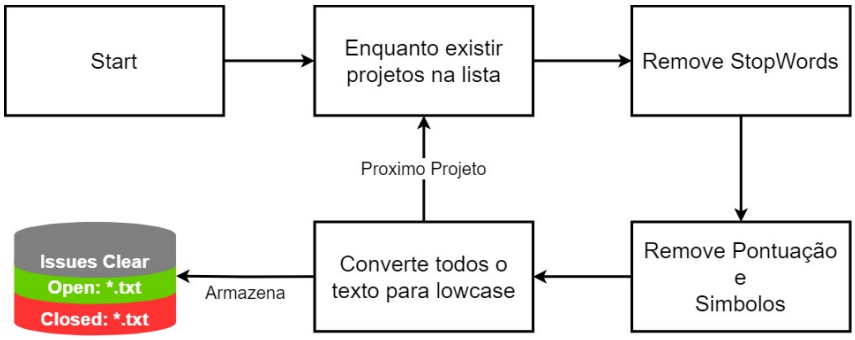

Figura 3.5: Processo de Tratamento de Dados.

Partindo para os processos de limpeza utilizando uma biblioteca em Python[79] chamada Natural Language Toolkit (NLTK), o algoritmo funcionou da seguinte forma: cada linha do arquivo *issue* era lido de maneira separada, depois essa sentença passa por uma tokenização e por fim aplica os processos de limpeza.

Primeiro foi feito a remoção dos *stop words* considerado um importante procedimento dentro do MD, onde são retirados dos dados um conjunto de palavras consideradas irrelevantes para o resultado, é importante lembrar que não existe uma lista universal para essas categorias de palavras, logo qualquer grupo pode ser escolhido conforme o objetivo alcançado, utilizando a biblioteca NLTK foi pego uma lista de *stop words* com uma série de termos em inglês como: *I, me, my, myself, ...*

A remoção de símbolos também recorreu ao NLTK para que caracteres como (*,-, ...) fossem retirados para se ter uma padronização dos dados além da conversão de todos os elementos textuais para minusculo facilitando as etapas posteriores.

3.4 Classificação de Dados

Após a conclusão do tratamento dos dados seção 3.3, nossa próxima etapa foi a classificação desses dados limpos.

Para que a classificação fosse feita de maneira mais abrangente entre os princípios da GDPR, LGPD e diretrizes éticas em IA, adotamos um modelo JSON de arquivo conforme a Figura 3.6 para as *issues open* e *closed*.

As palavras escolhidas foram baseadas nos princípios da GDPR [1], LGPD [2] e na literatura de diretrizes éticas em IA [3].

Por se tratar de um modelo de arquivo JSON é possível facilmente alterarmos qualquer que seja um desses parâmetros e assim obtivermos resultados diferentes nas próximas etapas.

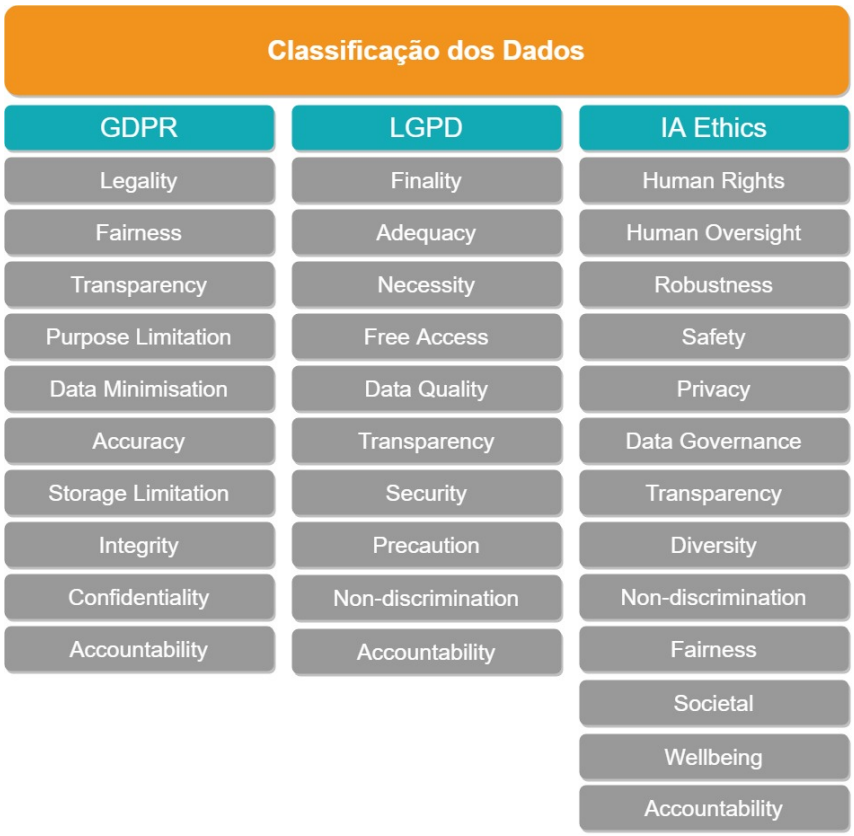

Figura 3.6: Palavras Escolhidas nos princípios da GDPR[1], LGPD [2] e Diretrizes Éticas IA [3] para Classificação de Dados.

Infelizmente os termos da LGPD tiveram quer ser traduzidos para que não houvesse exclusão por parte de alguns termos serem em português e grande maioria dos projetos serem em inglês. Outra questão está relacionada às limitações de usabilidade das ferramentas utilizadas na classificação desses dados, onde alguns princípios formados por palavras compostas tiveram que ser divididos em termos únicos por não ser possível o retorno de sinônimos de palavras compostas, sendo uma limitação imposta pela biblioteca NLTK.

Com nosso modelo de dados estabelecido, executamos um algoritmo conforme mostrado na Figura 3.7 onde é feito uma sequência de passos começando por *start*, para que as palavras chaves sejam extraídas de cada projeto e *issue* respectivamente.

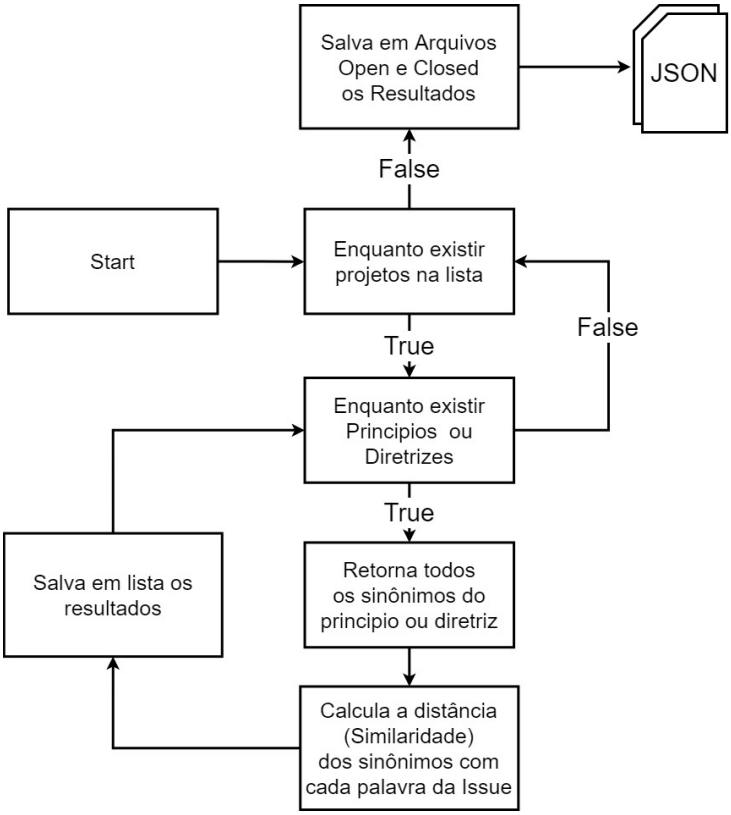

Figura 3.7: Processo de Classificação de Dados.

Enquanto a lista de projetos armazenada em nossa base seção 3.1 não for percorrida, cada princípio ou diretriz tem um levantamento de sinônimos realizado utilizando um parâmetro n, onde é estipulado um número de palavras próximas à aquele termo, em seguida uma biblioteca chamada *Difflib* é responsável por comparar sequências de dados em conjunto com a função *getCloseMatches* para pegar os elementos textuais das *issues* mais similares entre os sinônimos.

Para fazer uso desse recurso é necessário estabelecer um parâmetro *cutoff* antes da execução desse método, esse elemento restringe a porcentagem de semelhança entre palavras. Para esse trabalho foi escolhido um *cutoff=0,8*(80%) de similaridade entre os elementos textuais, resultando assim em termos encontrados nas *issues* desses projetos o mais próximo possível dos princípios ou diretrizes sem que resulte em termos sem sentido.

Com esses elementos extraídos, realizamos uma medida para saber com qual frequência essas palavras foram citadas de acordo com seus princípios ao longo de todas as *issues open* e *closed* dos projetos.

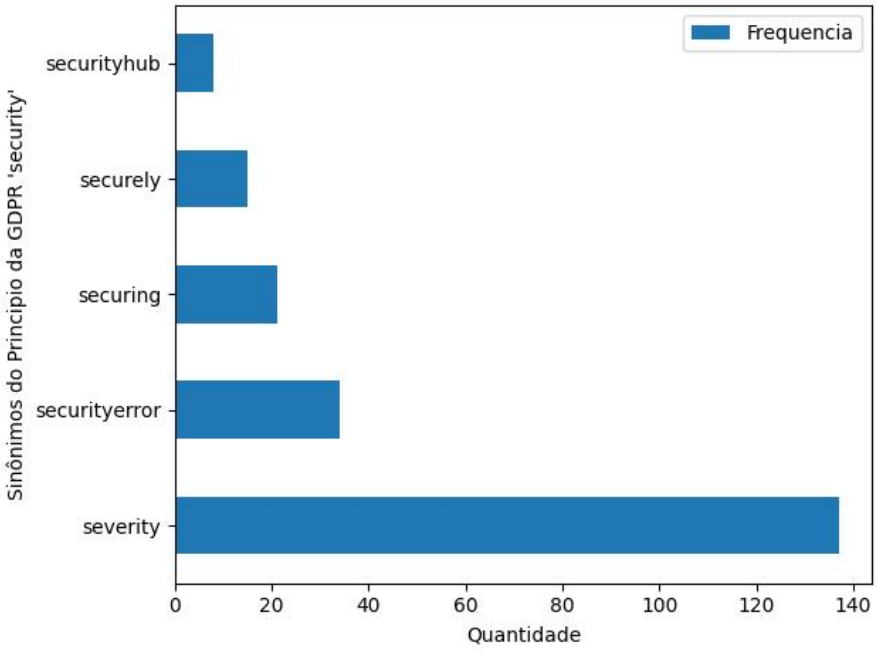

Figura 3.8: Gráfico de Sinônimos do Princípio (*Security*) da GDPR de Issues Open.

Gerando conforme o Gráfico 3.8 e 3.9 um exemplo usando um princípio para se ter uma noção dos resultados obtidos de tal maneira que o exemplo em questão faz um levantamento do princípio da GDPR de *Securitity* onde temos como resultados termos como (*Severity, Deprecation, Protection, ...*) mostrando assim a eficiência dos processos aplicados até agora.

Por fim geramos uma nuvem de palavras das *issues open* e *closed* dos repositórios conforme mostrado nas Figuras 3.10 e 3.11 para termos um apanhado geral de todas as palavras relacionadas ao GDPR, LGPD e ética IA obtidas ao final da nossa classificação. Analisando os Words Clouds conforme as Figuras 3.10 e 3.11 encontramos diversas palavras relacionadas com grande potencial de construção conforme a Tabela na seção 3.5 deste trabalho, mas também encontramos alguns *ruidos* que infelizmente não fazem sentido para nosso levantamento.

O objetivo por trás da criação dessa nuvem de palavras é apresentar palavras que poderiam entrar facilmente em nossa Tabela apresentada na seção 3.5, mas devido ao algoritmo que prioriza os temos com maior frequência acabam ficando de fora do resultado do trabalho.

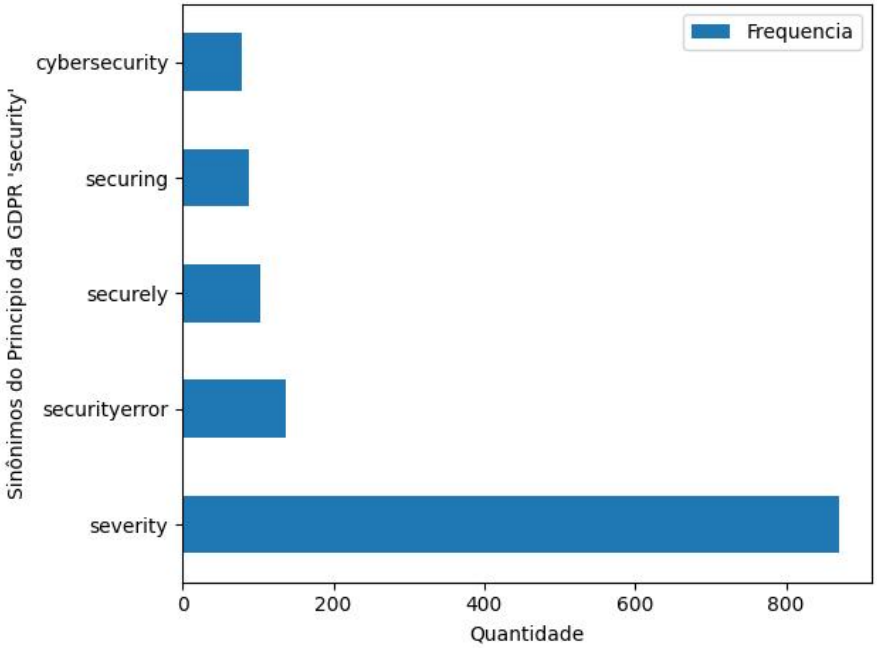

Figura 3.9: Gráfico de Sinônimos do Princípio (*Security*) da GDPR de Issues Closed.

Figura 3.10: Princípios GDPR, LGPD e Diretrizes Éticas IA da Issues Open dos Projetos.

Figura 3.11: Princípios GDPR, LGPD e Diretrizes Éticas IA da Issues Closed dos Projetos.

3.5 Geração da Tabela

Na etapa final da metodologia, partimos para a construção de forma automática da Tabela com as palavras-chave mais citadas em todos os projetos em determinado contexto como da GDPR, LGPD e ética IA conforme apresentando nas Figuras 3.4, 3.5 e 3.6.

Para ser gerado foi necessário que todas as etapas fossem concluídas com sucesso, por ser um processo em série ter os dados extraídos da fonte no *Data Extract* seção 3.2, realizar a limpeza na etapa de *Data Treatment* seção 3.3 e por fim a Classificação seção 3.4 para no fim chegar no objetivo principal deste trabalho, a geração da tabela.

Para a construção da tabela precisamos recorrer às frequências de palavras obtidas anteriormente na seção 3.4 em conjunto com o panda[81], um pacote Python[79] que fornece estruturas de dados rápidas, flexíveis e expressivas projetadas para tornar o trabalho com dados fácil e intuitivo.

Percorrendo a lista de frequências open e closed de cada regulação é feita uma remoção de termos idênticos ao princípio para garantir que não haja redundância entre eles, em seguida é realizado uma somatória de todas as palavras-chave de cada princípio para se ter uma noção de quantas vezes determinado principio foi citado nos projetos, por último conseguimos construir conforme as Tabelas 3.4, 3.6, 3.6 resultados de forma automatizada utilizando o recurso de *dataframe*.

De acordo com o apresentado, as Tabelas 3.4, 3.6, 3.6 foram geradas com sucesso mostrando a eficiência do algoritmo, após um grande período de operação para percorrer

Tabela 3.4: Resultado GDPR.

Princípios	Palavras-Chave	Total Citações
legality	reality, legally, illegality, egalite, legalipy	254
fairness	harness, equality, quit, flakiness,fitness	5049
transparency	transparent, transparently, transparencies, oil, tranparent, transparancy	15700
purpose limitation	proposed, purpose, limitations, invitation, limitation	3828
data minimisation	data, minification, cdata, minimization odata, datas	79617
accuracy	trust, truth, inaccuracy, truthy, trutv, strut	1657
storage limitation	storage, store, strange, stage, limitations, invitation	38459
integrity	unit, units, unify, junit, unity	18966
confidentiality	confidential, confidently, confidentially, identicality	96
accountability	separability, acceptability, unstability, venerability, vunerability, accountable, accountactivity	30

todos os 5250 projetos com grande número de *issues* disponibilizando uma variedade de termos os quais serão discutidos na Seção 4.

Tabela 3.5: Resultado LGPD.

Princípios	Palavras-Chave	Total Citações
finality	finally, inability, xfinity, conclusions, funtionality, funcionality, functinality	895
adequacy	efficiency, sufficient, sufficiently, insufficiently, adequate	966
necessity	unnecessary, requirements, requirement, necessary, unnecessarily	14585
free access	access, accessing, free, freeze, accessed	43411
data quality	data, quality, equality, quantity, inequality, cdata	84087
transparency	transparent, transparently, transparencies, oil, tranparent, transparancy	15700
security	certificate, sure, production, survey, certificates	29227
precaution	deprecation, location, creation, protection, projection	48363
non-discrimination	discrimination, discriminating, nondiscriminatedunion, discriminaton	57
accountability	separability, acceptability, unstability, venerability, vunerability, accountable, accountactivity	30

Tabela 3.6: Resultado IA.

Princípios	Palavras-Chave	Total Citações
human rights	right, human, rights, lights, bright	16006
human oversight	human, humans, huffman, oversights, oversight, overnight	813
robustness	business, harness, validity, sharness, hardline, hardness	1640
safety	safe, safely, unsafe, guard, guards	28458
privacy	exclusion, exclusions, securely, concealed, privates, secrecy	868
data governance	data, provenance, cdata, governance, odata, datas	79262
transparency	transparent, transparently, transparencies, oil, tranparent, transparancy	15700
diversity	arity, university, variety, diversify, divers, varient, divergences	622
non-discrimination	discrimination, discriminating, nondiscriminatedunion, discriminaton	57
fairness	harness, equality, quit, flakiness, fitness	5049
societal	social, soil, gnusocial, sociale, societe, socialite, socialspy	3946
wellbeing	beat, upbeat, opbeat, dwelling	294
accountability	separability, acceptability, unstability, venerability, vunerability, accountable, accountactivity	30

Capítulo 4

Resultados

Nesse capítulo foi discutido os dados obtidos na seção 3 e elaborados resultados conforme a construção das Tabelas 3.4, 3.5 e 3.6, além de analisar resultados paralelos encontrados posteriormente apos a conclusão do objetivo principal deste trabalho.

4.1 Análise de Erros, Tendências e Comportamento

Analisando os dados das Tabelas da seção 3.5, podemos extrair uma série de resultados paralelos ao nosso objetivo principal onde pode se observar o nível de ruído (*palavras sem contexto*) ao final do procedimento, analisar qual regulação é mais abordada pelos projetos, verificar se essas empresas já implementaram ou vão implementar essas *issues* regulatórias, esses são apenas algumas das interpretações escolhidas para este trabalho.

O ruído foi um dos problemas mais recorrentes desse projeto, porque esteve presente em todas as etapas da metodologia, desde a fase de extração de dados seção 3.2 com possíveis falhas, exemplo disso foi o excesso de velocidade entre os processos de código em comparação as requisições HTTP, resolvido com uso de wait() para esperar o carregamento total da requisição antes de extraí-la,

Outro tipo de ruído gerado pela etapa de tratamento de dados seção 3.3 onde não é possível obter um controle da lista de palavras consideradas stop words, mas o ruído que mais impactou os resultados deste trabalho está relacionado com a escolha dos parâmetros de classificação de dados seção 3.4, conforme apresentado nas Figuras 4.1 e 4.2.

Conforme a Figura 4.1 o parâmetro de similaridade usado foi de 75%, resultando em um grande ruído de palavras que enquanto estão próximas do princípio estruturalmente, sintaticamente são discrepantes como podemos ver em termos como *purple* resultado atrelado ao *purpose limitation* na GDPR.

Por outro lado, conforme a Figura 4.2 temos os seguintes resultados de classificação. As palavras encontradas possuem relação com os princípios, mas em compensação alguns

	GDPR
accountability	['mutability', 'accountant', 'runability', 'convertability', 'tractability', 'unstability', 'unability', 'notability', 261]
accuracy	['accurate', 'accurately', 'inaccuracy', 'accuray', 'accuracyop', 'accuracies', 'innacuracy', 1477]
confidentiality	['confidential', 'confident', 'confidently', 'coincidentally', 'conditionality', 'identicality', 'oneidentity', 69]
fairness	['failures', 'painless', 'fires', 'harness', 'flakiness', 12482]
integrity	['integrate', 'integer', 'inherit', 'intensity', 'eintegrity', 'integrityerror', 11947]
lawfulness	['flawless', 'fullness', 'lawful', 'helpfulness', 'lawless', 36]
minimisation	['animation', 'administration', 'initialisation', 'optimisation', 'minification', 9271]
purpose	['proposed', 'purposes', 'propose', 'prose', 'purple', 'proposer', 2066]
security	['severity', 'securityerror', 'securing', 'securely', 'securityhub', 'cybersecurity', 1493]
storage	['store', 'strange', 'stored', 'stage', 'stores', 28340]
transparency	['transparent', 'transparently', 'translucency', 'transparencies', 'tranparent', 'transparancy', 'transparant', 2627]

Figura 4.1: Resultado de Classificação da GDPR com Similaridade 75%.

	IA
accountability	[0]
data_governance	[0]
diversity	['variety', 95]
fairness	['equity', 7]
human_rights	[0]
non-discrimination	[0]
privacy	['secrecy', 'concealment', 'privateness', 'seclusion', 42]
robustness	['validity', 414]
safety	['safe', 'guard', 'rubber', 'condom', 'prophylactic', 'refuge', 5289]
societal	['social', 1925]
transparency	['trasparency', 'tranparency', 'foil', 'transparecy', 'transpareny', 'transparenc', 'transparence', 'transprency', 54]
wellbeing	['upbeat', 10]

Figura 4.2: Resultado de Classificação da IA com Similaridade 95%.

destes princípios da GDPR, LGPD e IA não obtiveram resultados como *accountability e data governance* podendo tornar as Tabelas da seção 3.5 incompletas.

Verificando a quantidade de citações por termos conforme as Tabelas 3.4, 3.5 e 3.6. Foi identificado através de uma quantização dessas citações que empresas deram mais atenção a determinados princípios pelo número de frequência que esses termos apareceram no resultado. Com esses valores foi possível estimar qual das regulações essas empresas estão alinhadas através do número total de referências de cada regulação nos projetos conforme apresentado na Tabela 4.1.

Tabela 4.1: Quantidade de Citações por Regulação.

Regulação/Diretriz	Total de Citações
GDPR	163.656
LGPD	237.321
Ética IA	152.745

Conforme os resultados apresentados na Tabela 4.1 foi verificado um alinhamento maior de projetos com a LGPD. Um dos motivos pode estar relacionado ao fato de alguns princípios extraídos da LGPD serem responsáveis por subirem esses números, por serem considerados termos genéricos muito utilizados no desenvolvimento como no caso da extração do princípio *data quality* onde a palavra-chave *data* alavancou esse resultado, podendo ser muitas vezes citada.

No contexto da LGPD, outra forma de interpretar esse resultado poderia ser que a LGPD é mais recente além de possui uma formulação de princípios voltados a realidade brasileira, ao contrário da GDPR que é mais antiga e consolidada no atual momento e tem como foco a União Europeia, logo foi preciso tomar cuidado ao analisar esses dados.

Usando os resultados das frequências das *issues open* e *closed* da seção 3.4, foi possível estimar com essas informações em que momento de implantação dessas regulações esses projetos se situam, usando os arquivos JSON de frequência com a quantidade de citações por categoria de issue foi criado conforme a Tabela 4.2 um conjunto de dados cujo resultado ajudou a compreender o atual momento do cenário de desenvolvimento que essas empresas se encontram.

Tabela 4.2: Quantidade de Citações por Categoria de Issue.

Regulação/Diretriz	Issue Open	Issue Closed
GDPR	33.605	130.051
LGPD	66.912	170.409
Ética IA	28.302	124.443
Total de Citações	122.739	424.903

De acordo com os resultados apresentados conforme a Tabela 4.2, mostrou que independente da regulação citada GDPR, LGPD ou ética IA , uma parte se encontra já implementada ao compararmos os números entre as categorias de *issue*.

De maneira generalizada entre os projetos é importante lembrar que alguns desses projetos podem não se enquadrar nessa situação, mas como se trata de resultados gerais uma parte se encontra nessa fase do desenvolvimento.

4.2 Verificação dos Resultados da Tabela com o Contexto das Regulações

Para validarmos os resultados apresentados nas Tabelas 3.4, 3.5, 3.6, foi feita uma pesquisa em alguns projetos de forma manual para saber se aquelas palavras-chave estavam associadas as *issues* e se estavam dentro de algum dos contextos abordados no trabalho GDPR, LGPD ou diretrizes éticas em IA.

Conforme apresentado na Tabela 4.3 foi encontrado relacionamentos entre a palavra *confidential* e uma issue sobre acesso indevido a dados confidenciais [82], que fazem parte do contexto de *confidentiality* conforme a Tabela 3.4.

Procurando pelos princípios da LGPD encontramos algumas *issues* relacionadas a transparência *transparency* como interface confusa para o usuário [83].

Olhando para as diretrizes éticas em IA podemos utilizar a governança de dados *data governance* que permite desenvolvedores propor novas soluções IA.

De forma geral, os resultados foram satisfatórios, pois uma grande parte das *issues* que foram encontradas nos projetos que tinham algo relacionado a palavra-chave analisada, realmente tinham relação com algum dos contextos abordados da GDPR, LGPD e ética em IA, mostrando que a metodologia obteve uma grande precisão no levantamento desses dados.

É importante citar também a limitação da pesquisa de verificação quando realizarmos buscas em projetos no GitHub® [71], temos como resultado algo similar aos elementos da busca que procuramos, logo nossa verificação pode ser comprometida em certos casos de busca.

Por se tratar de uma verificação manual os resultados encontrados das *issues* relacionadas as palavras-chave foram bem satisfatórios, pois mostrou uma boa precisão por parte da metodologia.

A partir de questões abordadas nos projetos que possuem as regulações da GDPR, LGPD ou diretriz ética IA mesmo que de forma indireta em seus títulos de *issues*, além de expor pequenas coisas que podem ser feitas para o aprimoramento da ferramenta

Tabela 4.3: Verificação Palavras-chave vs Issues.

Regulação	Principio	Palavra-chave	Issue
LGPD	necessity	requirements	odoo12: un checked "phone formatting"yet still my +XX is being removed [84]
LGPD	adequacy	adequate	sql: audit all information in the DB Console and ensure end points exist [85]
LGPD	finality	finally	[amp-consent] Fine grain, client-side consent [86]
LGPD	transparency	transparently	FSE: Content Slot block UI [83] is confusing to users
GDPR	accountability	accountable	Solicit feedback from users after pairing has finished [87]
GDPR	purpose limitation	purpose	Is there any user data that is either being collected or Tracked by the lottie-ios [88]
GDPR	confidentiality	confidential	A non-admin user is incorrectly able to see Cluster Settings from DB Console[82]
IA	safety	safely	Hashtag links are not URL safe [89]
IA	data governance	governance	Allow devs to specify that a new AI is just an idea [90]
IA	diversity	variety	AIE Model performing poorly if faces not varied enough? [91]

em trabalhos futuros como correção ortográfica de termos além de eliminação de termos redundantes.

4.3 Ameaças e Limitações para Verificação

Se tratando de ameaças externas para este trabalho o ponto que chama a atenção é a não possibilidade de garantir que todos os projetos analisados por nossa metodologia sejam de pequenas ou grandes empresas, logo, a quantidade de *issues* veio a oscilar no decorrer do processamento.

Como consequência a base de dados pode se tornar defasada para análises futuras por falta de elementos textuais para se analisar dentro desse contexto de ameaça externa temos a incapacidade de controlar a disponibilidade desses projetos por serem de terceiros.

Tratando das ameaças internas desse projeto, existe uma série de comportamentos e de tratamento de dados que poderiam ser evitados, sendo eles repetições de termos com nome idêntico, mas com erro de escrita por conta do desenvolvedor autor da issue, limitações das ferramentas de cálculo de similaridade com o tempo de processamento atrelado a complexidade dos algoritmos são considerados pontos de ameaça.

Por fim a escolha humana de alguns parâmetros para realizar a análises desses dados para garantir um bom balanceamento entre diversidade e quantidade de termos, sendo o

caso do parâmetro de porcentagem de semelhança entre o princípio e os elementos textuais de cada issue, onde é colocado de frente o seguinte impedimento, ao colocar um valor mais próximo de 100% temos palavras iguais ao princípio, mas em compensação perdemos os sinônimos dela, O contrário seria se reduzíssemos em valores inferiores o nível de ruído ou palavras que não possuem conexão seria elevado.

E por fim, ao tratar possíveis ameaças ao construto desse trabalho, podemos usar como exemplo a escolha por observação e testes dos parâmetros na classificação, mesmo que de certa forma essas escolhas estejam sustentadas com testes conforme as Figuras 4.1 e 4.2.

Além disso, as adaptações de alguns princípios para retornarem algo lógico, a falta de experiência em implementar e lidar com essas regulações da GDPR, LGPD e IA por parte do autor, são resultados de iterações testes, logo pode acontecer certos enganos na hora da interpretação e escolha de sinônimos para alguns princípios como foi o caso da LGPD.

Capítulo 5

Conclusão

Este trabalho demonstrou ser possível através da implementação de uma metodologia com uma série de etapas de *Extração, Tratamento, Classificação e Geração*, extrair de forma automatizada palavras-chave citadas no contexto das regulações da GDPR, LGPD e ética IA, que podem auxiliar desenvolvedores e outros profissionais na fase de implementação.

Ao observamos a frequência em que determinados termos foram citados nos repositórios, em conjunto com a nuvem de palavras gerada ao final da classificação de dados seção 3.4, é possível identificar quais elementos entre os demais possuem maior relevância entre os projetos estudados sendo eles: *purpose, equity, necessarily, ...*

Como resultado, a tabela gerada mostrou que podemos recorrer a técnicas e ferramentas da mineração de dados para extrair e compreender o comportamento dos diversos projetos de código aberto disponíveis na internet, para assim poder auxiliar futuros desenvolvedores na implementação das regulações da GDPR, LGPD e se for o caso ética em IA.

Nos resultados da GDPR se teve um foco maior em determinados princípios como: *data minimisation, storage limitation* e *integrity*, as palavras-chave que chamou atenção dentre esses termos foram termos como: *stage, proposed, minimization, data*, demostrando que alguns elementos são importantes no projeto da forma direta sem uso de alguns sinônimos.

No caso da *LGPD* termos como: *data quality, free access* e *precaution* foram os mais citados entre os projetos, demostrando que uma parte dos projetos tem certa preocupação com a qualidade desses dados em conjunto com o acesso livre, mas com um certo nível de precaução com o comportamento do projeto, além de mostrar que a LGPD está presente em soluções pelo mundo.

Por último a IA se preocupou com princípios como: *safety, data governance* e *transparency*, um comportamento esperado por esses projetos mostrando que essas soluções

estão preocupadas com essas questões como foi citado no início do trabalho onde tinha como objetivo garantir a ética e proteger os dados dos usuários.

De forma paralela, também podemos utilizar os resultados desse trabalho para entender e compreender o comportamento dos diversos projetos olhando apenas para as *issues closed* dessas soluções. É possível entender se esses projetos vinham em processo de implantação das regulações da GDPR, LGPD e ética IA, caso esse comportamento fosse comum nas *issues open*, infere que esses projetos poderiam estar na fase de início dessas implantações.

De forma geral, essa ferramenta está longe de ser perfeita, mas por agora oferece um resultado satisfatório a curto prazo, e a longo abre espaço para possíveis aprimoramentos e possibilidades para trabalhos futuros que podem surgir a respeito desse tema.

Para trabalhos futuros, pode ser implementadas novas funcionalidades para aumentar a precisão e exatidão da tabela automatizada, além de ferramentas para eliminar elementos indesejados, tais como, erro ortográfico extraído dos projetos, palavras sem conexão com o modelo de princípios e diretrizes, ou até fazer uso dessa metodologia para extrair informações fora do contexto da GDPR, LGPD ou ética em IA.

Referências

[1] PARLIAMENT, THE EUROPEAN e THE COUNCIL OF THE EUROPEAN UNION: *General data protection regulation (gdpr)*. Official Journal of the European Union, página 88, 2016. https://gdpr-info.eu/. ix, 7, 8, 23

[2] Republica, Presidente: *Lei geral de proteção de dados pessoais (lgpd)*. Diário Oficial [da] República Federativa do Brasil, 2019. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm. ix, 2, 3, 8, 9, 23

[3] AI (AI HLEG), High Level Expert Group on: *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. EUROPEAN COMMISSION, 2019. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. ix, 12, 23

[4] Tiwari, Saurabh e Santosh Singh Rathore: *A methodology for the selection of requirement elicitation techniques*, 2017. 1

[5] Borenstein, Jason e Ayanna M. Howard: *Emerging challenges in AI and the need for AI ethics education*. AI Ethics, 1(1):61–65, 2021. https://doi.org/10.1007/s43681-020-00002-7. 1

[6] Douglas, David M., David Howard e Justine Lacey: *Moral responsibility for computationally designed products*. AI Ethics, 1(3):273–281, 2021. https://doi.org/10.1007/s43681-020-00034-z. 1

[7] Writer, Christina Pazzanese Harvard Staff: *Ethical concerns mount as ai takes bigger decision-making role in more industries*. Harvard, 1(1), 2020. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/. 1

[8] Anderson, Susan Leigh e Michael Anderson: *AI and ethics*. AI Ethics, 1(1):27–31, 2021. https://doi.org/10.1007/s43681-020-00003-6. 1

[9] Anish, Preethu Rose, Balaji Balasubramaniam, Jane Cleland-Huang, Roel Wieringa, Maya Daneva e Smita Ghaisas: *Identifying architecturally significant functional requirements*. Em *2015 IEEE/ACM 5th International Workshop on the Twin Peaks of Requirements and Architecture*, páginas 3–8, 2015. 1

[10] Davey, Bill e Kevin R. Parker: *Requirements elicitation problems: A literature analysis*. Cornell University, 1:11, 2017. https://arxiv.org/pdf/1709.08481.pdf. 1

[11] Latif, Maria, Hira Shaheen, Taiba Wakeel e Tehseen Kousar: *Requirement elicitation process in software engineering.* International Journal of Scientific and Engineering Research, 10:8, 2019. https://www.ijser.org/researchpaper/Requirement-Elicitation-Process-in-Software-Engineering.pdf. 1

[12] Bani-Salameh, Hani e Nadera Al jawabreh: *Towards a comprehensive survey of the requirements elicitation process improvements.* Em *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*, IPAC '15, New York, NY, USA, 2015. Association for Computing Machinery, ISBN 9781450334587. https://doi.org/10.1145/2816839.2816872. 1

[13] Zhang, Zheying: *Effective requirements development - a comparison of requirements elicitation techniques.* Department of Computer Sciences University of Tampere, 1:9, 2007. https://www.researchgate.net/publication/228717829_Effective_Requirements_Development-A_Comparison_of_Requirements_Elicitation_Techniques. 1

[14] Salini, P. e S. Kanmani: *Elicitation of security requirements for e-health system by applying model oriented security requirements engineering (mosre) framework.* Em *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, CCSEIT '12, página 126–131, New York, NY, USA, 2012. Association for Computing Machinery, ISBN 9781450313100. https://doi.org/10.1145/2393216.2393238. 1, 12

[15] Sangaroonsilp, Pattaraporn, Hoa Khanh Dam, Morakot Choetkiertikul, Chaiyong Ragkhitwetsagul e Aditya Ghose: *A taxonomy for mining and classifying privacy requirements in issue reports.* CoRR, abs/2101.01298, 2021. https://arxiv.org/abs/2101.01298. 1, 15

[16] Ferrari, Alessio, Liping Zhao e Waad Alhoshan: *Nlp for requirements engineering: Tasks, techniques, tools, and technologies.* Em *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, páginas 322–323, 2021. 1

[17] Zhao, Liping, Waad Alhoshan, Alessio Ferrari, Keletso J. Letsholo, Muideen A. Ajagbe, Erol Valeriu Chioasca e Riza T. Batista-Navarro: *Natural language processing for requirements engineering: A systematic mapping study.* ACM Comput. Surv., 54(3), apr 2021, ISSN 0360-0300. https://doi.org/10.1145/3444689. 1, 2

[18] Sangaroonsilp, Pattaraporn, Hoa Khanh Dam, Morakot Choetkiertikul, Chaiyong Ragkhitwetsagul e Aditya Ghose: *A taxonomy for mining and classifying privacy requirements in issue reports.* Cornell University, 1:13, 2021. https://arxiv.org/abs/2101.01298. 1, 12

[19] Akay, Haluk e Sang Gook Kim: *Extracting functional requirements from design documentation using machine learning.* Procedia CIRP, 100:31–36, 2021, ISSN 2212-8271. https://www.sciencedirect.com/science/article/pii/S2212827121004637, 31st CIRP Design Conference 2021 (CIRP Design 2021). 1, 16

[20] *Three years of gdpr: the biggest fines so far.* https://www.bbc.com/news/technology-57011639. 2, 3

[21] Sauer, Andre Cunha e João Padilha Moreira: *Lgpd - lei geral de proteção de dados.* SEMINÁRIO DE TECNOLOGIA GESTÃO E EDUCAÇÃO, 3(2), nov. 2021. http://raam.alcidesmaya.edu.br/index.php/SGTE/article/view/354. 2

[22] Canedo, Edna Dias, Anderson Jefferson Cerqueira, Rogério Machado Gravina, Vanessa Coelho Ribeiro, Renato Camões, Vinicius Eloy dos Reis, Fábio Lúcio Lopes de Mendonça e Rafael T. de Sousa Jr.: *Proposal of an implementation process for the brazilian general data protection law (LGPD).* Em Filipe, Joaquim, Michal Smialek, Alexander Brodsky e Slimane Hammoudi (editores): *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*, páginas 19–30. SCITEPRESS, 2021. https://doi.org/10.5220/0010398200190030. 2

[23] Martins, Antônio Diogo Forte, Patrícia Vieira da Silva Barros, José Maria Monteiro e Javam de Castro Machado: *LGPD: A formal concept analysis and its evaluation.* Em *Anais do XXXV Simpósio Brasileiro de Bancos de Dados, SBBD 2020, online, September 28 - October 1, 2020*, páginas 259–264. SBC, 2020. https://doi.org/10.5753/sbbd.2020.13651. 2, 3

[24] Cerqueira, José Antonio Siqueira de: *Exploring ethical requirements elicitation for applications in the context of ai.* Repositório UNB, 2021. https://repositorio.unb.br/handle/10482/41966. 2, 12

[25] Gold, Nicolas E. e Jens Krinke: *Ethics in the mining of software repositories.* Empir. Softw. Eng., 27(1):17, 2022. https://doi.org/10.1007/s10664-021-10057-7. 2, 15

[26] Zhang, Wei, Wei Wei, Lingjie Xu, Lingling Jin e Cheng Li: *AI matrix: A deep learning benchmark for alibaba data centers.* CoRR, abs/1909.10562, 2019. http://arxiv.org/abs/1909.10562. 2

[27] Steck, Harald, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond e Justin Basilico: *Deep learning for recommender systems: A netflix case study.* AI Mag., 42(3):7–18, 2021. https://doi.org/10.1609/aimag.v42i3.18140. 2

[28] Choi, Tae Rang e Minette E. Drumwright: *"ok, google, why do I use you?" motivations, post-consumption evaluations, and perceptions of voice AI assistants.* Telematics Informatics, 62:101628, 2021. https://doi.org/10.1016/j.tele.2021.101628. 2

[29] Milossi, Maria, Eugenia Alexandropoulou Egyptiadou e Konstantinos E. Psannis: *AI ethics: Algorithmic determinism or self-determination? the GPDR approach.* IEEE Access, 9:58455–58466, 2021. https://doi.org/10.1109/ACCESS.2021.3072782. 2, 3

[30] Baker-Brunnbauer, Josef: *Management perspective of ethics in artificial intelligence.* AI Ethics, 1(2):173–181, 2021. https://doi.org/10.1007/s43681-020-00022-3. 2

[31] Cormack, Andrew: *Thinking with GDPR: A guide to better system design.* Inf. Serv. Use, 41(1-2):61–69, 2021. https://doi.org/10.3233/ISU-210107. 3

[32] Canedo, Edna Dias, Angelica Toffano Seidel Calazans, Anderson Jefferson Cerqueira, Pedro Henrique Teixeira Costa e Eloisa Toffano Seidel Masson: *Agile teams' perception in privacy requirements elicitation: Lgpd's compliance in brazil.* Em *2021 IEEE 29th International Requirements Engineering Conference (RE)*, páginas 58–69, 2021. 3

[33] Alhazmi, Abdulrahman e Nalin Asanka Gamagedara Arachchilage: *I'm all ears! listening to software developers on putting GDPR principles into software development practice.* Pers. Ubiquitous Comput., 25(5):879–892, 2021. https://doi.org/10.1007/s00779-021-01544-1. 3

[34] Biable, Seblewongel Esseynew, Nuno Manuel Garcia, Dida Midekso e Nuno Pombo: *Ethical issues in software requirements engineering.* Software, 1(1):31–52, 2022, ISSN 2674-113X. https://www.mdpi.com/2674-113X/1/1/3. 3, 12

[35] Dias Canedo, Edna e Bruno Cordeiro Mendes: *Software requirements classification using machine learning algorithms.* Entropy, 22(9), 2020, ISSN 1099-4300. https://www.mdpi.com/1099-4300/22/9/1057. 4

[36] Jindal, Rajni, Ruchika Malhotra, Abha Jain e Ankita Bansal: *Mining non-functional requirements using machine learning techniques.* e Informatica Softw. Eng. J., 15(1), 2021. https://doi.org/10.37190/e-inf210105. 4

[37] Supervisor, European Data Protection: *The history of the general data protection regulation.* Official Journal of the European Union, página 88, 2018. https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en. 7

[38] PINHEIRO e PATRICIA P.: *PROTEÇÃO DE DADOS PESSOAIS: COMENTÁRIOS À LEI N. 13.709/2018 (LGPD).* Editora Saraiva, 2021, ISBN 9786555595123. 8, 10

[39] LIMA, Ana Paula Moraes Canto D.: *LGPD Aplicada.* Grupo GEN, 2021, ISBN 9788597026931. 9

[40] Oliveira, Ricardo: *LGPD: Como evitar as sanções administrativas.* Editora Saraiva, 2021, ISBN 9786553623262. 9, 11

[41] LIMA e Cíntia Rosa Pereira D.: *ANPD e LGPD: Desafios e perspectivas.* Grupo Almedina, 2021, ISBN 9786556272764. 9

[42] GARCIA, Lara R: *Lei Geral de Proteção de Dados (LGPD): Guia de implantação.* Editora Blucher, 2020, ISBN 9786555060164. 10

[43] Republica, Presidente da: *ConstituiÇÃo da repÚblica federativa do brasil de 1988*. Diário Oficial [da] República Federativa do Brasil, 1988. http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. 10

[44] Ferrão, Sâmmara Éllen Renner, Artur Potiguara Carvalho, Edna Dias Canedo, Alana Paula Barbosa Mota, Pedro Henrique Teixeira Costa e Anderson Jefferson Cerqueira: *Diagnostic of data processing by brazilian organizations - A low compliance issue*. Inf., 12(4):168, 2021. 10

[45] Faceli, K.: *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC, 2011, ISBN 9788521618805. https://integrada.minhabiblioteca.com.br/#/books/978-85-216-2146-1. 11

[46] Coppin, B. e J.D.P. Valério: *Inteligência artificial*. Grupo Gen - LTC, 2015, ISBN 9788521629351. https://books.google.com.br/books?id=ZEK2AQAACAAJ. 11

[47] Galaz, Victor, Miguel A. Centeno, Peter W. Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum, Darryl Farber, Joern Fischer, David Garcia, Timon McPhearson, Daniel Jimenez, Brian King, Paul Larcey e Karen Levy: *Artificial intelligence, systemic risks, and sustainability*. Technology in Society, 67:101741, 2021, ISSN 0160-791X. https://www.sciencedirect.com/science/article/pii/S0160791X21002165. 11, 12

[48] Miernicki, Martin e Irene Ng (Huang Ying): *Artificial intelligence and moral rights*. AI Soc., 36(1):319–329, 2021. https://doi.org/10.1007/s00146-020-01027-6. 12

[49] Aguilar, Alfonso Robles, Jorge Octavio Ocharán-Hernández e Ángel Juan Sánchez-García: *A systematic mapping study of artificial intelligence in software requirements*. Res. Comput. Sci., 149(11):179–188, 2020. http://rcs.cic.ipn.mx/2020_149_11/A%20Systematic%20Mapping%20Study%20of%20Artificial%20Intelligence%20in%20Software%20Requirements.pdf. 12

[50] Chen, Yi e Yun Li: *Intelligent autonomous pollination for future farming - A micro air vehicle conceptual framework with artificial intelligence and human-in-the-loop*. IEEE Access, 7:119706–119717, 2019. https://doi.org/10.1109/ACCESS.2019.2937171. 12

[51] Chukkapalli, Sai Sree Laya, Sudip Mittal, Maanak Gupta, Mahmoud Abdelsalam, Anupam Joshi, Ravi S. Sandhu e Karuna P. Joshi: *Ontologies and artificial intelligence systems for the cooperative smart farming ecosystem*. IEEE Access, 8:164045–164064, 2020. https://doi.org/10.1109/ACCESS.2020.3022763. 12

[52] Wang, Zeyu, Mingyu Li, Jia Lu e Xin Cheng: *Business innovation based on artificial intelligence and blockchain technology*. Inf. Process. Manag., 59(1):102759, 2022. https://doi.org/10.1016/j.ipm.2021.102759. 12

[53] Grassi, Alessia e Mauro Vallati: *An exploratory study on the use of artificial intelligence to initiate legal understanding for business development.* FLAP, 8(4):1065–1082, 2021. https://collegepublications.co.uk/ifcolog/?00046. 12

[54] Bostrom, Nick e Eliezer Yudkowsky: *The ethics of artificial intelligence*, página 316–334. Cambridge University Press, 2014. 12

[55] IEEE: *ETHICALLY ALIGNED DESIGN - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html. 12

[56] R., Goldschmidt, Emmanuel Passos e Eduardo Bezerra: *Data Mining*, volume 2. Elsevier Editora, 2015, ISBN 978-85-352-7822-4. https://integrada.minhabiblioteca.com.br/books/9788595156395/, Acesso em: 2021 ago. 22. 13, 14

[57] Wu, Dan e Yuanjun Shen: *English feature recognition based on GA-BP neural network algorithm and data mining.* Comput. Intell. Neurosci., 2021:1890120:1–1890120:10, 2021. https://doi.org/10.1155/2021/1890120. 13, 14

[58] Huai, Yangting e Qianxiao Zhang: *Construction of social security fund cloud audit platform based on fuzzy data mining algorithm.* Complex., 2021:9939454:1–9939454:11, 2021. https://doi.org/10.1155/2021/9939454. 13, 14

[59] Li, Wei, Qiling Zhou, Junying Ren e Samantha Spector: *Data mining optimization model for financial management information system based on improved genetic algorithm.* Inf. Syst. E Bus. Manag., 18(4):747–765, 2020. https://doi.org/10.1007/s10257-018-00394-4. 13, 14

[60] Wu, You, Zheng Wang e Shengqi Wang: *Human resource allocation based on fuzzy data mining algorithm.* Complex., 2021:9489114:1–9489114:11, 2021. https://doi.org/10.1155/2021/9489114. 14

[61] Bagui, Sikha e Richard P. Stanley: *Mining frequent itemsets from streaming transaction data using genetic algorithms.* J. Big Data, 7(1):54, 2020. https://doi.org/10.1186/s40537-020-00330-9. 14

[62] Qiao, Mengke e Kewei Huang: *Correcting misclassification bias in regression models with variables generated via data mining.* Inf. Syst. Res., 32(2):462–480, 2021. https://doi.org/10.1287/isre.2020.0977. 14

[63] Chien, Jen-Tzung: *Deep bayesian data mining.* Em Caverlee, James, Xia (Ben) Hu, Mounia Lalmas e Wei Wang (editores): *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, páginas 865–868. ACM, 2020. https://doi.org/10.1145/3336191.3371870. 14

[64] Abeysinghe, Rashmie e Licong Cui: *Query-constraint-based mining of association rules for exploratory analysis of clinical datasets in the national sleep research resource.* BMC Medical Informatics Decis. Mak., 18(S-2):89–100, 2018. https://doi.org/10.1186/s12911-018-0633-7. 14

[65] Javed, Muhammad Fasih, Waqas Nawaz e Kifayat-Ullah Khan: *HOVA-FPPM: flexible periodic pattern mining in time series databases using hashed occurrence vectors and apriori approach.* Sci. Program., 2021:8841188:1–8841188:14, 2021. https://doi.org/10.1155/2021/8841188. 14

[66] Abualigah, Laith Mohammad e Akram Jamal Dulaimi: *A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm.* Clust. Comput., 24(3):2161–2176, 2021. https://doi.org/10.1007/s10586-021-03254-y. 14

[67] Zhou, YiTao: *Natural language processing with improved deep learning neural networks.* Sci. Program., 2022:6028693:1–6028693:8, 2022. https://doi.org/10.1155/2022/6028693. 15

[68] Vydiswaran, V.G.Vinod, Xinyan Zhao e Deahan Yu: *Data science and natural language processing to extract information in clinical domain.* Em *5th Joint International Conference on Data Science and Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD 2022, página 352–353, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 9781450385824. https://doi.org/10.1145/3493700.3493773. 15, 16

[69] Omar, Marwan, Soohyeon Choi, DaeHun Nyang e David Mohaisen: *Robust natural language processing: Recent advances, challenges, and future directions.* CoRR, abs/2201.00768, 2022. https://arxiv.org/abs/2201.00768. 15

[70] Raharjana, Indra Kharisma, Daniel Siahaan e Chastine Fatichah: *User stories and natural language processing: A systematic literature review.* IEEE Access, 9:53811–53826, 2021. 15, 16

[71] Preston-Werner, Tom, Chris Wanstrath, Scott Chacon e P. J. Hyett: *Where the world builds software*, 2008. https://github.com/. 18, 19, 33

[72] *Giving back to open source for the holidays.* https://24pullrequests.com/. 19

[73] *I want to get involved!* https://up-for-grabs.net/. 19

[74] Schneeman, Richard: *Help out your favorite open source projects and become a better developer while doing it.* https://www.codetriage.com/. 19, 20

[75] *I want to get involved!* https://firstcontributions.github.io/. 19

[76] *beginner-friendly.* https://github.com/topics/beginner-friendly. 19

[77] *easy.* https://github.com/topics/easy. 19

[78] *first-timers-only.* https://github.com/topics/first-timers-only. 19

[79] Rossum, Guido van: *Python is a programming language that lets you work quickly and integrate systems more effectively.* https://www.python.org/. 20, 22, 27

[80] *Beautiful soup is a python library for pulling data out of html and xml files.* https://www.crummy.com/software/BeautifulSoup/bs4/doc/. 20

[81] *Powerful data structures for data analysis, time series, and statistics.* https://pypi.org/project/pandas/. 27

[82] awoods187: *A non-admin user is incorrectly able to see cluster settings from db console.* https://github.com/cockroachdb/cockroach/issues/74692. 33, 34

[83] shaunandrews: *Fse: Content slot block ui is confusing to users.* https://github.com/Automattic/wp-calypso/issues/34889. 33, 34

[84] outsourcing f1: *odoo12: un checked "phone formatting" yet still my +xx is being removed.* https://github.com/odoo/odoo/issues/62227. 34

[85] awoods187: *sql: audit all information in the db console and ensure end points exist.* https://github.com/cockroachdb/cockroach/issues/57763. 34

[86] mcmd: *[amp-consent] fine grain, client-side consent.* https://github.com/ampproject/amphtml/issues/26735. 34

[87] tansaku: *Solicit feedback from users after pairing has finished.* https://github.com/AgileVentures/WebsiteOne/issues/1215. 34

[88] venkata m: *Is there any user data that is either being collected or tracked by the lottie-ios.* https://github.com/airbnb/lottie-ios/issues/1262. 34

[89] roschaefer: *[bug] hashtag links are not url safe.* https://github.com/Human-Connection/Human-Connection/issues/208. 34

[90] GuiltyDolphin: *Allow devs to specify that a new ia is just an idea.* https://github.com/duckduckgo/community-platform/issues/1243. 34

[91] Kirin-kun: *Iae model performing poorly if faces not varied enough?* https://github.com/deepfakes/faceswap/issues/283. 34