UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS

DEPARTAMENTO DE COMPUTAÇÃO

LEXI DOS SANTOS MOREIRA DE OLIVEIRA

Orientador: Prof. Dr. Guilherme Tavares de Assis

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA PARA GERAÇÃO SEMIAUTOMÁTICA E DINÂMICA DE DADOS ESTATÍSTICOS RELATIVOS À VIOLÊNCIA CONTRA A POPULAÇÃO LGBTQIA+ BRASILEIRA

Ouro Preto, MG

2023

UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

LEXI DOS SANTOS MOREIRA DE OLIVEIRA

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA PARA GERAÇÃO SEMIAUTOMÁTICA E DINÂMICA DE DADOS ESTATÍSTICOS RELATIVOS À VIOLÊNCIA CONTRA A POPULAÇÃO LGBTQIA+ BRASILEIRA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto - UFOP como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Guilherme Tavares de Assis

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Lexi dos Santos Moreira de Oliveira

Proposta e desenvolvimento de uma estratégia para geração automática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 25 de Agosto de 2023.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
João Fernando dos Santos Vilela (Examinador) - Bacharel - Psicólogo
Isabelle Mayumi Koga (Examinadora) - Bacharel - Teach Lead - Efí

Guilherme Tavares de Assis, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 25/08/2023.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/08/2023, às 12:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0577025** e o código CRC **F97CDBCD**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho a minha mãe, quem ensinou-me o significado de força e a delicadeza do amor

Agradecimentos

Agradeço a minha mãe, que foi minha guia e me deu forças para que eu pudesse alcançar meus sonhos e objetivos.

Agradeço ao meu orientador, Guilherme Tavares de Assis, por todo apoio, paciência e por acreditar em meu potencial, foi essencial para a elaboração deste trabalho.

Agradeço aos meus amigos que estiveram presentes comigo neste processo durante a graduação, aprendi muito com vocês, obrigada por terem tornado meus dias mais leves.

Agradeço a Mirelly, por todo apoio e incentivo que proporcionou em minha trajetória, pelos momentos que vivenciamos juntas e pelo afeto que sempre encontrei ao estar com você.

Agradeço a Cibele, pelos bons momentos que compartilhamos juntas e por estar ao meu lado nos desafios que a graduação apresentou.

Agradeço a Maria Eduarda, Leidiane e Lara, pelo apoio, por serem minha família em Ouro Preto e por tornarem o final da minha graduação um dos momentos mais felizes de minha vida.

Por fim, em especial, agradeço a Tácita e Natália, irmãs que a vida me proporcionou e que tornaram maravilhoso todo o processo de concretização deste sonho.

"Querer ser livre é também querer livres os outros."

(Simone de Beauvoir, escritora).

Resumo

No Brasil, a população LGBTQIA+ enfrenta variadas formas de violência e são poucas as fontes de dados que evidenciam essa realidade, o que dificulta o entendimento da dimensão desta e a criação de ações efetivas para combatê-la. Tendo como base este problema, o presente trabalho propôs, desenvolveu e validou uma primeira versão funcional de uma estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira. A estratégia é composta pela coleta de páginas contendo informações sobre a violência contra LGBTQIA+ brasileiros, onde dados relevantes são extraídos automaticamente destas páginas, armazenados e integrados e, a partir deles, dados estatísticos são gerados e apresentados. Pretende-se, com esta estratégia, contribuir com uma fonte confiável de informação sobre a violência enfrentada por LGBTQIA+ brasileiros que auxilie a compreender este problema e possa ser utilizada como base para ações que visem mitigá-lo. De forma geral, a primeira versão desta estratégia apresentou-se promissora, já que sua precisão média geral é de 44%, considerando experimentos com os processos de coleta, extração e integração de dados. A versão inicial da estratégia mostrou que esta pode ser uma ferramenta útil para geração semiautomática e dinâmica de dados e evidências sobre a violência contra a população LGBTQIA+ brasileira. Este trabalho também pretende contribuir para o conhecimento e a prática do uso de técnicas de coleta, extração, integração, análise e visualização de dados relacionados à violência contra população LGBTQIA+ brasileira.

Palavras-chave: Geração semiautomática de dados estatísticos. Violência contra LGBTQIA+. Coleta semiautomática de dados. Extração automática de dados. Integração de dados. Visualização de dados.

Abstract

In Brazil, the LGBTQIA+ population faces various forms of violence, and there are few sources of data that highlight this reality, which hinders the understanding of the extent of this issue and the creation of effective actions to combat it. Building upon this problem, this study proposed, developed, and validated an initial functional version of a strategy for the semi-automatic and dynamic generation of statistical data related to violence against the brazilian LGBTQIA+ population. The strategy comprises the collection of pages containing information about violence against LGBTQIA+ individuals in Brazil, where relevant data is automatically extracted from these pages, stored, integrated, and used to generate and present statistical information. The aim of this strategy is to contribute to a reliable source of information about the violence faced by LGBTQIA+ individuals in Brazil, aiding in understanding this problem and serving as a foundation for actions aimed at mitigating it. Overall, the initial version of this strategy showed promise, with an overall average accuracy of 44%, considering experiments with data collection, extraction, and integration processes. The initial version of the strategy demonstrated that it can be a useful tool for the semi-automatic and dynamic generation of data and evidence about violence against the brazilian LGBTQIA+ population. This work also aims to contribute to the knowledge and practice of using techniques for data collection, extraction, integration, analysis, and visualization related to violence against the brazilian LGBTQIA+ population.

Keywords: Semi-automatic generation of statistical data. Violence against LGBTQIA+. Semi-automatic data collection. Automatic data extraction. Data integration. Data visualization.

Lista de Ilustrações

Figura 2.1 – Violências cometidas contra a população LGBTQIA+ brasileira no segundo semestre de 2022 segundo a ONDH 7
Figura 2.2 – Vítimas de violência que são LGBTQIA+ 8
Figura 2.3 – Dados dos Assassinatos de pessoas transgênero no Brasil entre 2008 e 2021 8
Figura 2.4 – Quantitativo de mortes violentas de LGBT+, Brasil, entre 1963-2021 9
Figura 2.5 – Exemplo de arquivo XML 13
Figura 3.1 – Arquitetura de funcionamento da estratégia proposta para geração de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira . 20
Figura 3.2 – Combinações entre perfis LGBTQIA+ e termos de violência 23
Figura 3.3 – XML resultante do processo completo de extração a partir de uma notícia coletada 27
Figura 3.4 – Escolha do arquivo contendo perfis LGBTQIA+ 36
Figura 3.5 – Seleção dos perfis LGBTQIA+ 37
Figura 3.6 – Seleção do arquivo contendo termos de violência 37
Figura 3.7 – Seleção dos termos de violência 37
Figura 3.8 – Seleção do arquivo contendo domínios de sites de notícias 38
Figura 3.9 – Seleção dos domínios de sites de notícias 38
Figura 3.10–Especificação da quantidade de notícias a serem retornadas 38
Figura 3.11–Links de páginas de notícias encontradas pelo coletor 39
Figura 3.12–Tela inicial da plataforma *Qlik Cloud* 40
Figura 3.13–Tela de criação de uma nova aplicação 40
Figura 3.14–Tela *upload* de dados para a aplicação 41
Figura 3.15–Tela *upload* de arquivo para a aplicação 41
Figura 3.16–Tela de dados obtidos a partir de um arquivo de entrada 42
Figura 3.17–Conjuntos criados pela aplicação a partir do XML de entrada 42
Figura 3.18–Conexão entre os conjuntos de dados criados 43
Figura 3.19–Tela de sucesso do carregamento dos dados 43
Figura 3.20–Tela de seleção da opção para criação de visualizações de dados 44
Figura 3.21–Painel principal para criação de visualizações de dados 45
Figura 3.22–Exemplo de gráfico criado a partir dos dados de violência por região do Brasil 45
Figura 3.23–Visualização de dados com filtro ativo 46
Figura 3.24–Informações presentes em partes dos gráficos gerados 46
Figura 3.25–Exemplo de gráfico gerado por um algoritmo próprio 47
Figura 4.1 – Índices de casos de violência contra LGBTQIA+ brasileiros em relação aos anos 59
Figura 4.2 – Faixa etária das vítimas de casos de violência contra LGBTQIA+ brasileiros 60

Figura 4.3 – Índices sobre a orientação sexual das vítimas de casos de violência contra LGBTQIA+ brasileiros 60
Figura 4.4 – Índices sobre a identidade de gênero das vítimas de casos de violência contra LGBTQIA+ brasileiros 61
Figura 4.5 – Índices de violência contra LGBTQIA+ nas regiões do Brasil 62
Figura 4.6 – Índices sobre os tipos de violências mais frequentes contra LGBTQIA+ brasileiros 62

Lista de Tabelas

Tabela 2.1 – Diferença entre dados estruturados e semi-estruturados 13
Tabela 4.1 – Perfis LGBTQIA+ utilizados na busca por notícias de violência contra LGBTQIA+ 49
Tabela 4.2 – Termos de violência utilizados na busca por notícias de violência contra LGBTQIA+ 49
Tabela 4.3 – Domínios dos sites de notícias utilizados na busca por notícias de violência contra LGBTQIA+ 49
Tabela 4.4 – Tabela com os resultados dos experimentos de coleta realizados 52
Tabela 4.5 – Tabela com os resultados dos experimentos de extração de dados realizados 53
Tabela 4.6 – Tabela com os resultados dos experimentos de integração de dados realizados 58

Lista de Abreviaturas e Siglas

ANTRA	Associação Nacional de Travestis e Transexuais
BI	*Business Intelligence*
CNFE	Chinese News Fact Extractor
EAC	Extração Automática de Conteúdo
GGB	Grupo Gay da Bahia
IA	Inteligência Artificial
LGBTQIA+	Lésbicas, Gays, Bissexuais, Transgêneros, Travestis, *Queers*, Intersexuais, Assexuais e mais
MDHC	Ministério dos Direitos Humanos e da Cidadania
OMS	Organização Mundial da Saúde
ONDH	Ouvidoria Nacional de Direitos Humanos
XML	*Extensible Markup Language*

Sumário

1 Introdução 1
1.1 Justificativa 1
1.2 Objetivos Geral e Específicos 2
1.3 Método do Trabalho 3
1.4 Organização do Trabalho 3
2 Revisão de Literatura 4
2.1 Fundamentação Teórica 4
2.1.1 População LGBTQIA+ e seus Contextos 4
2.1.2 Violência Contra a População LGBTQIA+ no Brasil 6
2.1.3 Coletas de páginas na *Web* 9
2.1.4 Extração de Dados da *Web* 10
2.1.5 Armazenamento de Dados Semi-estruturados 12
2.1.6 Visualização de Dados 14
2.2 Trabalhos Relacionados 15
3 Desenvolvimento 19
3.1 Arquitetura de funcionamento da estratégia proposta 19
3.1.1 Coleta semiautomática de páginas 21
3.1.2 Extração de dados 25
3.1.2.1 Extração da data de publicação da notícia 28
3.1.2.2 Extração do nome da vítima 29
3.1.2.3 Extração da idade da vítima 30
3.1.2.4 Extração da identidade de gênero e orientação sexual da vítima 30
3.1.2.5 Extração da localidade em que a vítima sofreu violência . . . 32
3.1.2.6 Extração dos tipos de violência sofridos pela vítima 34
3.1.3 Integração de dados 35
3.2 Interfaces da ferramenta resultante da estratégia proposta 36
3.2.1 Interface da coleta de páginas de notícias sobre violência contra LGBTQIA+ brasileiros 36
3.2.2 Interfaces da plataforma de visualização de dados *Qlik Cloud* 39
4 Experimentação Prática 48
4.1 Métrica de Avaliação 48
4.2 Descrição dos Experimentos 48
4.3 Análise dos Resultados Obtidos 51
4.3.1 Coleta semiautomática de páginas *Web* contendo notícias sobre violências contra LGBTQIA+ brasileiros 52
4.3.2 Extração dos dados 52

4.3.3 Integração dos dados 57
4.3.4 Visualização dos dados obtidos 58
5 Considerações Finais 64
5.1 Conclusão 64
5.2 Trabalhos Futuros 64

Referências 65

1 Introdução

Atualmente, a violência contra a população Lésbicas, Gays, Bissexuais, Travestis, Transgêneros, *Queers*, Intersexuais, Assexuais e mais (LGBTQIA+) é uma realidade inaceitável, ainda vivenciada em todo mundo, ocasionada pelo simples fato de que as pessoas que compõem esta população divergem de normativas e padrões sociais impostos sobre seus corpos. A violência contra LGBTQIA+ no Brasil é um problema sistêmico e histórico herdado do colonialismo. Para GOMES et al. (2021), é notável que no Brasil o preconceito representa os esteriótipos idealizados pelas crenças enraizadas na sociedade e, segundo MARTINS (2018), a identidade moral e cultural brasileira "remonta à colonização europeia e aos seus valores cristãos, brancos, ocidentais, misóginos e heterossexuais" implicando em uma sociedade que, desde o Brasil colônia, repreende o que destoa da moralidade conservadora do colonizador. De acordo com GOMES et al. (2021), a população LGBTQIA+ é a mais afetada pela discriminação por suas identidade de gênero e orientação sexual destoarem das socialmente normalizadas, tornando-as mais suscetíveis a diversos tipos de violências.

As violências enfrentadas pela população LGBTQIA+ brasileira vão desde a negação de direitos básicos, como um lar e educação, até a mesmo a assassinatos brutais e hediondos ocasionados por ódio a seus corpos e vivências. OLIVEIRA e MOTT (2022) apresentam dados coletados pelo Grupo Gay da Bahia (GGB) em 2021, destacando o Brasil como o país onde mais LGBTQIA+ são assassinados: estima-se uma morte a cada 29 horas. Os autores também exibem que desde o ano de 1980 foram contabilizadas mais de seis mil mortes de LGBTQIA+ brasileiros, sendo que mais de três mil delas ocorreram entre os anos de 2010 e 2019.

A violência no Brasil contra a população LGBTQIA+ é um problema complexo, influenciado por fatores religiosos, culturais, sociais e também pela falta de proteção legal adequada. Dada sua complexidade, é necessário que propostas de soluções sejam multifacetadas. Legislações, políticas públicas, programas de prevenção e combate à violência podem ajudar a mitigar o problema mas, para que tais possibilidades de solução existam, é preciso entender, por meio de dados confiáveis e precisos, a realidade da violência enfrentada pela população LGBTQIA+ brasileira.

1.1 Justificativa

A escassez de dados confiáveis e precisos sobre a violência contra a população LGBTQIA+ brasileira é um enorme desafio para a compreensão desta realidade e o combate a este problema tão grave. As estatísticas existentes, divulgadas por fontes distintas como canais de ouvidoria de Direitos Humanos do Governo Federal e por instituições não governamentais como a Associação Nacional de Travestis e Transexuais (ANTRA) e o GGB, muitas vezes são subnotificadas ou

incompletas devido a diversos fatores, tais como possíveis falhas na coleta manual desses dados e casos de violência não denunciados ou registrados inadequadamente.

Neste contexto, a semiautomatização da coleta e a automatização da apresentação de dados significativos sobre a violência contra LGBTQIA+ no Brasil, reunidas em uma única estratégia ou ferramenta, podem constituir um passo importante para o combate a esta violência, pois ajudaria na compreensão desta realidade explicitando-a com os dados obtidos. A utilização de técnicas computacionais para coleta, extração, armazenamento e visualização desses dados pode tornar o processo mais rápido, eficiente e eficaz do que as abordagens utilizadas atualmente.

Embora a automação possa ser efetiva e promissora, ainda há pouca pesquisa e prática sobre como coletar, extrair e analisar dados sobre a violência contra a população LGBTQIA+ brasileira. É necessário, portanto, investigar e desenvolver estratégias de geração e apresentação automatizadas de dados estatísticos que sejam precisas, confiáveis e relevantes para a compreensão e prevenção da violência contra essa população.

1.2 Objetivos Geral e Específicos

Este presente trabalho tem, como objetivo geral, a proposta, o desenvolvimento e validação da primeira versão completa e funcional de uma estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira. A pesquisa busca, portanto, contribuir para o avanço do conhecimento e da prática no uso de técnicas de coleta, extração, análise e apresentação de dados relacionados à violência contra essa população, com intento de influenciar políticas públicas e ações efetivas de combate à violência e promoção dos direitos humanos.

De maneira geral, os objetivos específicos alcançados neste trabalho são:

- geração de dados e evidências que podem servir como fonte confiável de informação para formuladores de políticas públicas, ativistas e organizações que trabalham para a prevenção e redução da violência contra a população LGBTQIA+ no Brasil;
- compreensão da extensão da violência contra a população LGBTQIA+ no Brasil por meio dos dados estatísticos gerados;
- contribuição, por meio da apresentação e visualização de dados estatísticos de maneira clara e acessível, para o aumento da conscientização sobre a prevalência e gravidade da violência contra a população LGBTQIA+ brasileira;
- contribuição para promoção da justiça social, dos direitos humanos e da igualdade para a população LGBTQIA+ no Brasil;
- geração de um repositório de páginas da Web contendo notícias sobre violência contra LGBTQIA+ brasileiros a partir de processos de coleta semiautomática;

- geração de um banco de dados semi-estruturado contendo dados obtidos a partir de notícias sobre violência contra LGBTQIA+ brasileiros.

1.3 Método do Trabalho

Para alcançar o objetivo geral descrito, foi definida e elaborada uma arquitetura que descreve o funcionamento de uma estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira.

Para validar tal arquitetura, uma versão inicial de todos os seus componentes foi implementada e validada experimentalmente de forma prática. Estes experimentos envolveram coleta semiautomática de páginas *Web*, extração de dados, integração de dados, geração e visualização de dados estatísticos para avaliar a precisão da estratégia proposta.

1.4 Organização do Trabalho

O restante deste trabalho está estruturado como se segue. No Capítulo 2, é realizada uma revisão de literatura abordando a fundamentação teórica e trabalhos relacionados ao tema proposto. No Capítulo 3, é descrita de forma detalhada a estratégia proposta para geração semiautomática e dinâmica de dados estatísticos sobre violência contra a população LGBTQIA+ brasileira. No Capítulo 4, são discutidos os experimentos, os resultados e as interfaces da estratégia proposta. Por fim, no Capítulo 5, são apresentadas as conclusões obtidas a partir do presente trabalho e as perspectivas de trabalho futuro.

2 Revisão de Literatura

Este capítulo aborda a revisão de literatura que serve de base para o presente trabalho. Para atingir tal objetivo, organiza-se em duas seções: Seção 2.1, que explora a fundamentação teórica de suporte ao desenvolvimento, e Seção 2.2, que expõe os trabalhos diretamente relacionados.

2.1 Fundamentação Teórica

Nesta seção, é contemplada a fundamentação teórica necessária para o entendimento e realização adequada deste trabalho. A Subseção 2.1.1 especifica e contextualiza população LGBTQIA+. A Subseção 2.1.2 apresenta dados sobre violência cometida contra a população LGBTQIA+ brasileira. A Subseção 2.1.3 aborda coleta de páginas na *Web*. A Subseção 2.1.4 retrata a extração de dados da *Web*. A Subseção 2.1.5 refere-se a dados semi-estruturados e suas possíveis formas de armazenamento. E, por fim, a 2.1.6 aborda ferramentas para visualização de dados.

2.1.1 População LGBTQIA+ e seus Contextos

Os seres humanos são, em sua natureza, plurais e distintos de inúmeras formas. Apesar da clara diferença entre cada indivíduo, através da história da humanidade regras e normativas foram estabelecidas como padrões a serem seguidos no contexto do convívio social. Entre estas, há as que tentam estabelecer um padrão sobre a orientação sexual e a identidade de gênero das pessoas. De acordo com BUTLER (2003, p. 59) "o gênero é a estilização repetida do corpo, um conjunto de atos repetidos no interior de uma estrutura reguladora altamente rígida, a qual se cristaliza no tempo para produzir a aparência de uma substância, de uma classe natural de ser".

Neste aspecto, pode-se observar o gênero como uma construção social que permeia e regula as relações sociais humanas. No Brasil, há dois gêneros que são socialmente aceitos e tidos como naturais: homem e mulher, considerados os binários. O gênero é imposto como natural nas condições em que pessoas nascidas do sexo biológico masculino devem identificar-se e performar enquanto homens; o mesmo acontece com pessoas do sexo biológico feminino, que devem reconhecer-se enquanto mulheres.

A identidade de gênero de uma pessoa dá-se por suas vivências particulares, onde a mesma reconhece a qual gênero se sente pertencente. Sendo o gênero algo socialmente imposto a partir de características corporais, há pessoas que não se enquadram no gênero designado ao nascimento. Estas são as pessoas transgênero, que diferente das cisgênero, pessoas que se identificam com o gênero atribuído ao nascimento, não se conformam com as normas criadas pela sociedade sobre seus corpos.

Ir contra essa normativa regulamentadora do gênero faz com que essas pessoas sofram as mais diversas represálias, opressões e violências por parte da sociedade. Pessoas transgênero perdem desde o acesso à educação e saúde, estes básicos para a vivência humana, até mesmo a vida, esta tirada por pessoas que não aceitam a existência e vivência de seus corpos. No contexto da sigla e população LGBTQIA+, as pessoas transgênero e travestis, identidade de gênero específica brasileira, enquadram-se na letra T.

Enquanto pessoas transgênero, há também as pessoas não-binárias, aquelas que não se enquadram no binarismo de gênero, não se identificando apenas enquanto homem ou mulher. Não-binário é um termo que agrupa de forma genérica uma quantidade diversa de identidades de gênero que não se enquadram no binário homem-mulher. O não-binário pode ser representado tanto pela letra T, quanto pela letra Q da sigla LGBTQIA+, onde o Q representa a palavra *queer*, palavra esta de origem da língua inglesa utilizada para representar de forma ampla as identidades de gênero e sexualidades que destoam das normativas existentes.

Segundo OLIVEIRA (2010 apud APA, 2008):

> A orientação sexual refere-se ao indivíduo como alguém que tem uma identidade pessoal e social com base nas suas atrações, manifestando determinados comportamentos e aderindo a uma comunidade de pessoas que compartilham da mesma orientação sexual.

Atualmente, a orientação sexual é classificada sobre um olhar social em três formas distintas: heterossexualidade, homossexualidade e bissexualidade. Pessoas heterossexuais são aquelas que se relacionam afetiva e sexualmente por pessoas do gênero oposto dentro do binário de gênero: homens que se relacionam com mulheres, e mulheres que se relacionam com homens, sendo estes cisgêneros ou transgêneros. Por homossexualidade, compreende-se a orientação sexual de pessoas que se relacionam com outras pessoas do mesmo gênero: homens que se relacionam com homens e mulheres que se relacionam com mulheres. Por bissexualidade, compreende-se a orientação sexual em que o indivíduo relaciona-se com pessoas de qualquer gênero. Por fim, há também a assexualidade, conceito que vem ganhando mais espaço e sendo mais reconhecido, contemplando pessoas que sentem pouca ou nenhuma atração sexual independentemente de gênero.

No contexto brasileiro, a heterossexualidade é a orientação sexual socialmente normalizada. As pessoas, que destoam desse padrão, sofreram e ainda sofrem também diversas represálias e violências. Apenas em 1990, a Organização Mundial da Saúde (OMS) retirou a homossexualidade da CID-10, desclassificando a mesma como doença. O casamento entre pessoas homossexuais no Brasil só foi reconhecido como legal em 2013, por uma resolução publicada pelo Conselho Nacional de Justiça. E, no momento, ainda há uma série de direitos que não contemplam pessoas que não são heterossexuais apenas por sua orientação sexual.

As pessoas que se reconhecem enquanto lésbicas, gays, bissexuais e assexuais se enquadram nas letras L, G, B e A da sigla LGBTQIA+. As definições feitas neste trabalho são

de caráter informativo sobre os maiores grupos que compõem a população LGBTQIA+, mas, como essa é plural e complexa, alguns grupos não foram citados e encontram-se abrangidos pelo símbolo + da sigla. Entenda-se como população LGBTQIA+ como sendo um coletivo de pessoas que não se enquadram nas normas de gênero e orientação sexual socialmente estabelecidas.

2.1.2 Violência Contra a População LGBTQIA+ no Brasil

No cenário brasileiro, a população LGBTQIA+ ainda é vítima de muitos preconceitos e violências; particularmente, segundo a Associação Nacional de Travestis e Transexuais (ANTRA), o Brasil é o país com o maior número de assassinato de pessoas transgênero (BENEVIDES, 2022). A violência contra a população LGBTQIA+ vai desde sua educação, negada por seus pertencentes não serem aceitos na escola ou em seu próprio lar, sendo muitos forçados a procurarem uma independência ainda jovens, até a falta de oportunidades no mercado de trabalho, onde muitos são obrigados a sobreviverem em subempregos e são marginalizados por não terem seus direitos garantidos. A violência também é institucional, segundo BULGARELLI et al. (2021 apud MINAYO; MOREIRA et al., 2006, 2020):

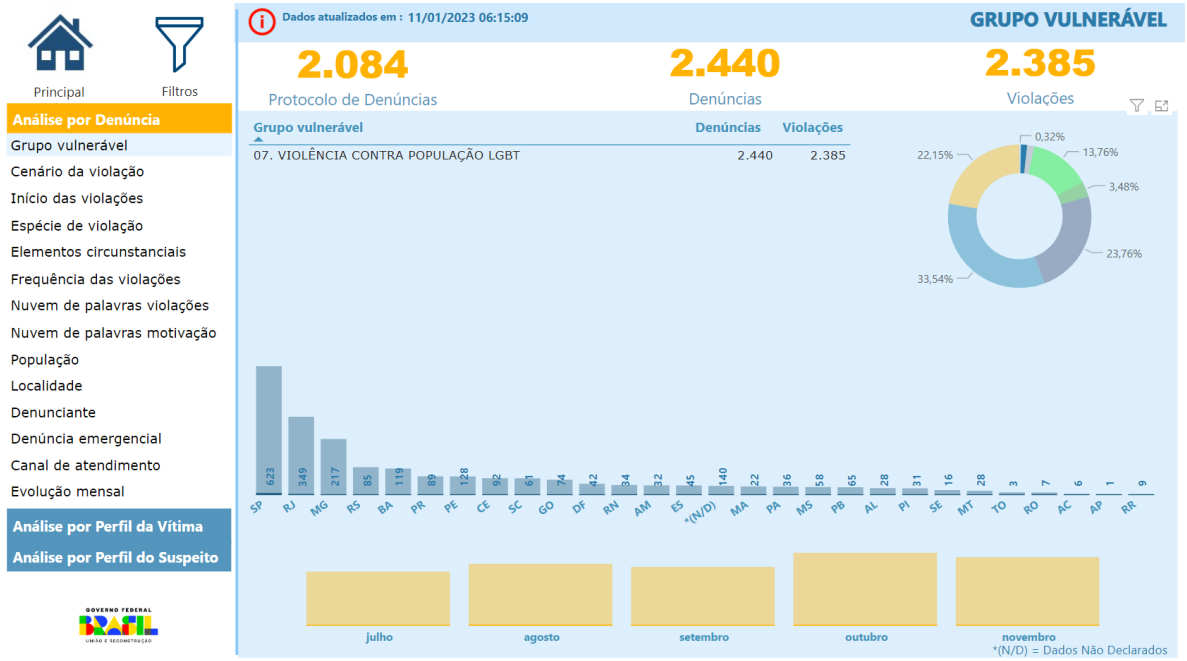
> Entendemos como violência institucional LGBTIfóbica toda forma de violência praticada por um agente ou órgão público que dificulte ou prejudique o acesso da vítima LGBTI aos mecanismos de justiça e ao reconhecimento integral da violência sofrida.

Ainda não há atualmente, por parte do Estado, uma fonte de dados que apresente quais são as características e especificidades da população LGBTQIA+ no Brasil. Desta forma, ainda torna-se difícil a criação de políticas públicas que visem possibilitar uma vida mais digna a essa população, já que não há muitos dados que exibam e retratem sua realidade.

Alguns dados sobre violência contra a população LGBTQIA+ brasileira podem ser encontrados na base de dados geradas pelo Disque Direitos Humanos (Disque 100), que possui relatórios de dados de denúncias com registro de violações de direitos humanos de 2011 a 2022. No site do Ministério dos Direitos Humanos e da Cidadania (MDHC), existe o Painel de Dados da Ouvidoria Nacional de Direitos Humanos (ONDH) que detalha dados colhidos pelos canais Disque 100, Disque 180 e o aplicativo Direitos Humanos Brasil. Por meio deste Painel, é possível acompanhar os dados de denúncias de violações de direitos humanos que ocorrem no Brasil e chegam ao conhecimento da ONDH.

O Painel de Dados da ONDH permite a utilização de filtros para a visualização de dados e, por meio destes, é possível identificar e categorizar denúncias. Um exemplo é, por meio da opção Análise por Denúncia, filtrar pelo grupo vulnerável composto pelos casos de violência contra a população LGBT (termo também utilizado para descrever a população LGBTQIA+), como retratado na Figura 2.1.

Figura 2.1 – Violências cometidas contra a população LGBTQIA+ brasileira no segundo semestre de 2022 segundo a ONDH



Fonte: Retirada do Painel da ONDH no site do MDHC.

Na Figura 2.1, pode-se visualizar a quantidade de Protocolos de Denúncias[1], Denúncias[2] e Violações [3] registradas no segundo semestre de 2022 já filtradas para o grupo vulnerável descrito. É possível visualizar também a quantidade de denúncias ocorridas em cada estado em todo o semestre ou em um mês específico e a porcentagem de denúncias registradas para esse grupo em relação a todas as denúncias registradas.

Outra opção que o painel oferece é a Análise por Perfil da Vítima, que consiste em filtrar o perfil de vítimas de denúncias registradas levando em consideração características das mesmas como, por exemplo, profissão, faixa etária, nacionalidade, etnia etc., sem levar em consideração a motivação da violência. Um dos perfis incluídos na filtragem é o LGBT, que exibe os dados de vítimas de violência que são pertencentes a comunidade LGBTQIA+. A filtragem por LGBT na Análise por Perfil da Vítima pode ser conferida na Figura 2.2.

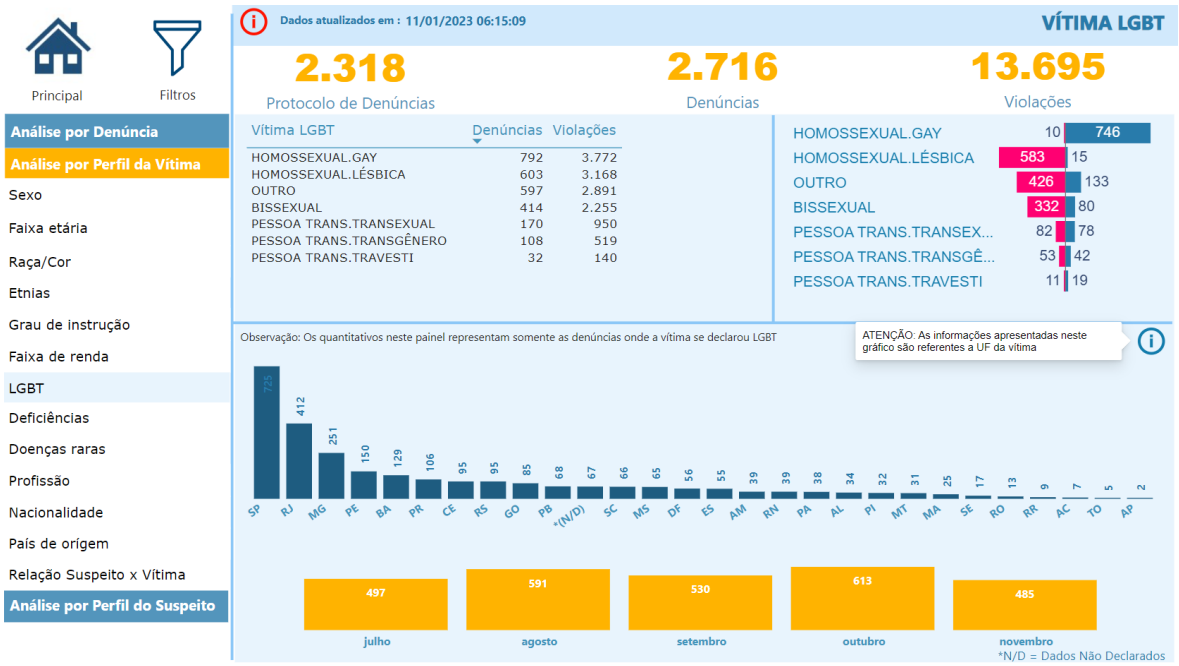
Como se pode perceber, o gráfico exibe também a contabilização de Protocolos de Denúncias, Denúncias e Violações acometidas contra vítimas com o perfil citado. As vítimas contabilizadas são identificadas como: Homossexual.Gay, Homossexual.Lésbica, Outro, Bissexual, Pessoa Trans.Transexual, Pessoa Trans.Transgênero e Pessoa Trans.Travesti. Para cada uma destas identidades, também tem-se um gráfico contabilizando as vítimas pelo seu sexo biológico. Também é possível visualizar a quantidade de vítimas por estado dentro de intervalo específico,

[1] Quantidade de registros que demonstra a quantidade de vezes em que os usuários buscaram a ONDH para registrarem uma denúncia. Um protocolo de denúncia pode conter uma ou mais denúncias.

[2] Quantidade de relatos de violação de direitos humanos envolvendo uma vítima e um suspeito. Uma denúncia pode conter uma ou mais violações de direitos humanos.

[3] Qualquer fato que atente ou viole os direitos humanos de uma vítima.

Figura 2.2 – Vítimas de violência que são LGBTQIA+



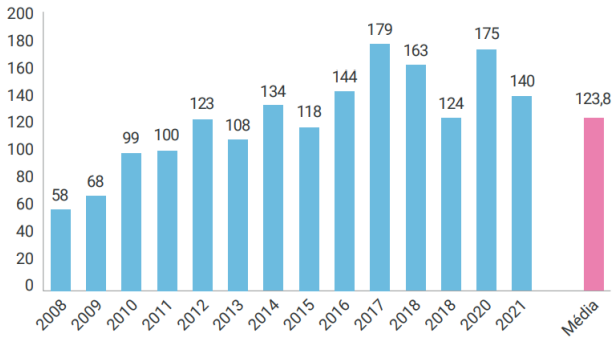
Fonte: Retirada do Painel da ONDH no site do MDHC.

sendo ele um ano, semestre ou mês.

Há também os dossiês e relatórios gerados pela ANTRA e o GGB que coletam informações de notícias sobre violência e divulgam em seus relatórios anuais. É por meio desses dados que é gerada uma visão mais ampla da violência contra as pessoas LGBTQIA+ brasileiras.

No dossiê divulgado pela ANTRA em 2022, referente aos dados coletados em 2021, é possível perceber dados de violências acometidas contra pessoas transgêneros no Brasil entre os anos de 2008 e 2021, como se pode observar na Figura 2.3.

Figura 2.3 – Dados dos Assassinatos de pessoas transgênero no Brasil entre 2008 e 2021



Fonte: Dossiê Assassinatos e Violências contra Travestis e Transexuais Brasileiras em 2021 organizado pela ANTRA.

A Figura 2.3 apresenta os dados de pessoas transgênero assassinadas entre 2008 e 2021, exibindo a quantidade de assassinatos contabilizados em cada ano e a média de mortes desses anos contabilizados. Observa-se a repetição de duas colunas do gráfico referentes ao ano de 2018,

o que se pode supor ser um erro de digitação e umas das colunas referir-se aos dados de 2019, dada a ordem crescente dos anos citados no gráfico.

O GGB possui, em seu Relatório 2021, uma tabela (vide Figura 2.4) que informa o quantitativo de mortes violentas de pessoas LGBTQIA+ brasileiras.

Figura 2.4 – Quantitativo de mortes violentas de LGBT+, Brasil, entre 1963-2021

Período	N. Vítimas
1963-1969	30
1970-1979	41
1980-1989	369
1990-1999	1.256
2000-2009	1.429
2010-2019	3.029
2020-2021	537
Total	**6.691**

Fonte: Mortes Violentas de LGBT+ no Brasil - Relatório 2021 organizado pelo GGB

A tabela apresenta o quantitativo de mortes violentas de pessoas LGBTQIA+ ocorridas no Brasil entre os anos de 1963 e 2021, exibindo períodos distintos entre estes anos e as vítimas contabilizadas em cada período. Esses dados têm origem em relatórios realizados pelo GGB desde seu início.

Como já mencionado, este trabalho gera, de uma forma semiautomatizada e dinâmica, dados estatísticos sobre violência contra a população LGBTQIA+ brasileira, contribuindo para o entendimento da realidade das violências sofridas por LGBTQIA+ brasileiros e podendo auxiliar, por exemplo, na criação de políticas públicas que visem diminuir as violências acometidas contra esta população, já que existem dados retratando as mesmas.

2.1.3 Coletas de páginas na *Web*

Com o conhecimento humano em constante expansão, continuamente aumenta-se a quantidade de informações disponíveis na *Web* e, consequentemente, também o número de páginas *Web*. Nesse contexto, fazem-se necessários mecanismos que possibilitem a busca efetiva e eficaz de páginas relacionadas aos conteúdos pesquisados por um usuário da Internet. Existem, atualmente, máquinas de buscas disponíveis que retornam páginas a partir de termos informados; um exemplo bem conhecido de uma delas é o *Google*.

A máquina de busca da *Google*, frequentemente referido apenas como "*Google*", é uma ferramenta fundamental para a recuperação de informações na internet. Ela utiliza algoritmos sofisticados e técnicas de indexação para fornecer resultados de busca altamente relevantes aos

usuários. O coração da máquina de busca da *Google* reside em seus algoritmos de indexação e classificação. Conforme explicado por Brin e Page (1998), o *Google* utiliza o *PageRank*, um algoritmo de análise de *links*, para avaliar a importância das páginas da *Web* com base no número e qualidade dos *links* que apontam para elas. Esse algoritmo forma a base para a classificação dos resultados de busca, garantindo que páginas mais prioritárias e relevantes apareçam no topo ao serem retornadas para o usuário.

Um recurso fornecido pelo *Google* são os parâmetros de busca avançada. Estes parâmetros oferecem aos usuários uma maneira mais refinada e precisa de direcionar suas pesquisas. Entre esses parâmetros, existe, por exemplo, o recurso de ”domínios de site”, que permite aos usuários restringir os resultados de pesquisa a um domínio específico, como ”.edu” para instituições educacionais ou ”.gov” para sites governamentais. Isso ajuda a filtrar os resultados e concentrar a busca em fontes confiáveis e relevantes. Ao inserir ”site:” seguido do domínio desejado e termos de pesquisa, os resultados serão limitados a páginas contidas apenas nesse domínio. Essa funcionalidade é útil, por exemplo, para pesquisas acadêmicas ou profissionais, pois permite aos usuários focalizar sua pesquisa em locais específicos da *Web*, aumentando a precisão e a eficácia da recuperação de informações desejadas.

Diferentemente de máquinas de busca de propósito geral que fazem uma varredura por toda a *Web*, como o *Google*, existem os coletores *Web* temáticos, ou simplesmente, coletor temático. Um coletor temático varre apenas a parte da *Web* que considera relevante para as necessidades do usuário. Isso resulta em uma varredura mais eficiente e direcionada, com uma maior probabilidade de encontrar informações relevantes. Alguns coletores temáticos podem obter automaticamente informações de diferentes fontes, como sites de notícias, redes sociais, blogs, entre outros. Segundo ASSIS (2008, p. iii):

> “[...] coletores temáticos servem para gerar coleções de páginas menores e restritas, já que apresentam o propósito maior de coletar páginas que sejam, da melhor forma possível, relevantes a um tópico ou interesse específico do usuário, a partir de uma especificação mais detalhada do que se deseja coletar.”

Este trabalho utiliza o *Google* para coleta de páginas de notícias contendo casos de violência contra LGBTQIA+ brasileiros. Combinações de termos que indicam violência contra LGBTQIA+, em conjunto com domínios de sites de notícias são fornecidos para uma busca avançada por casos de violência contra LGBTQIA+ em sites brasileiros de notícias.

2.1.4 Extração de Dados da *Web*

Navegar pela *Web* permite entrar em contato com uma vasta quantidade de informações. Em diversos cenários, tais como quando se deseja obter dados de sites e os mesmos não possuem formas de acesso a seus dados ou uma API oficial, faz-se necessário extrair essas informações disponíveis em páginas *Web* e o desafio encontra-se em extraí-las adequadamente. Neste aspecto,

existem dois problemas gerais: extrair informações de textos de linguagem natural e extrair informações estruturadas de páginas da *Web* (LIU, 2011, p. 363).

Alguns padrões de reconhecimento e tratamento podem ser gerados para recuperação de informações de textos em forma de linguagem natural; neste sentido, alguns trabalhos como o de LOPES et al. (2009), que envolve a análise e identificação de termos compostos (conceitos expressos em mais de uma palavra) considerando o domínio da Medicina, e o de PAPINENI et al. (2002), que qualifica o processo de tradução automática utilizando de recuperação de informações e interpretação de texto, ambos envolvendo Processamento de Linguagem Natural, têm sido desenvolvidos. Há formas também de, a partir da estrutura da página *Web* como por exemplo seus elementos HTML, realizar a extração de dados. O processo de extração permite agrupar e comparar dados de diversas fontes distintas.

Segundo LIU (2011, p. 363), pesquisadores e empresas de *internet* começaram a trabalhar no problema de extração no meio dos anos 90 e existem três abordagens principais:

a **Abordagem manual:** observando uma página *Web* e seu código fonte, o programador humano encontra alguns padrões e então escreve um programa para extrair o dado almejado. Para tornar o processo simples para programadores, várias linguagens de especificação de padrão e interfaces de usuários têm sido criadas. No entanto, essa abordagem não é escalável para um grande número de sites.

b ***Wrapper Induction*:** a partir de uma coleção de páginas rotuladas manualmente ou de registros de dados, um conjunto de regras de extração é aprendido. Esse aprendizado é realizado por meio do processo de aprendizagem supervisionada e é semiautomático. As regras são então empregadas para extrair itens de dados de outras páginas formatadas de forma semelhante.

c **Extração automática:** por meio do fornecimento de várias ou uma única página, automaticamente encontra padrões ou gramáticas para a extração de dados por meio de aprendizagem não supervisionada[4], eliminando o esforço da rotulagem manual para a extração e possibilitando escalar a extração de dados para um grande número de sites e páginas *Web*.

Uma das abordagens para a extração de dados a partir de uma fonte textual é a mineração de texto, conhecida também como *Text Mining*. Segundo RIBEIRO et al. (2015), esta abordagem consiste em um processo responsável por identificar padrões de conhecimento em documentos utilizando fontes textuais não estruturadas e "[...] técnicas de mineração de texto possibilitam

[4] De acordo com MONARD e BARANAUSKAS (2003) o algoritmo de *aprendizagem não-supervisionado* "analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira" e, após a definição dos agrupamentos, geralmente realiza uma análise "para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado".

uma análise de alto nível das características dos textos, permitindo uma análise qualitativa e quantitativa sobre o conteúdo de um ou mais documentos.".

Este trabalho utiliza de mineração de texto, combinada a abordagens manuais, para obtenção de dados sobre a violência contra a população LGBTQIA+ brasileira a partir de notícias presentes em páginas *Web*.

2.1.5 Armazenamento de Dados Semi-estruturados

A forma como os dados organizam-se na *Web* não é padronizada; há formas de organização diversas e distintas. Alguns exemplos de dados são imagens, vídeos, textos, planilhas e bases de dados, que são cotidianamente consumidos da *Web*. Esses dados podem ser classificados em três formas distintas (EBERENDU, 2016): estruturados, semi-estruturados e não estruturados.

Dados estruturados podem ser definidos como “os dados que possuem formato definido e comprimento, fácil de armazenar e analisar com alto grau de organização” (SIMÕES, 2022 apud EBERENDU, 2016, p. 14)

Já os dados semi-estruturados, de acordo com SIMÕES (2022 apud EBERENDU, 2016, p. 15) consistem em aqueles que possuem um pouco mais de estrutura do que comparados aos dados não estruturados; eles possuem mais flexibilidade, podendo mudar rapidamente, mas não seguem um esquema fixo. Segundo MELLO et al. (s.d. apud BUNEMAN, 1997, p. 2), também é possível afirmar que dados semi-estruturados:

> "[...] são dados nos quais o esquema de representação está presente (de forma explícita ou implícita) juntamente com o dado, ou seja, o mesmo é auto-descritivo. Isto significa que uma análise do dado deve ser feita para que a sua estrutura possa ser identificada e extraída."

Por fim, dados não estruturados, segundo SIMÕES (2022, p. 16):

> “[...] não possuem uma estrutura definida relacionada a modelos ou esquemas de dados predefinidos [...]” e “[...] são categorizados como dados qualitativos, ou seja não podem ser processados usando ferramentas convencionais e métodos, e não podem ser organizados em banco de dados relacionais, ao invés disso podem ser gerenciados em banco de dados não relacionais.”

A Tabela 2.1 mostra algumas distinções entre dados estruturados e semi-estruturados.

Dadas as distinções entre os dois tipos de dados, tem-se também a diferença entre como esses dados são armazenados e consultados em banco de dados. Enquanto os dados estruturados podem ser armazenados, por exemplo, em bancos de dados relacionais[5], já que possuem seus tipos e estruturas especificados, os dados semi-estruturados seguem uma abordagem diferente

[5] Segundo MACÁRIO e BALDO (2005) um banco de dados relacional ”é um conjunto de uma ou mais relações com nomes distintos. O esquema do banco de dados relacional é a coleção dos esquemas de cada relação que compõe o banco de dados.”.

Tabela 2.1 – Diferença entre dados estruturados e semi-estruturados

Dados estruturados	Dados semi-estruturados
Esquema predefinido	Nem sempre há um esquema predefinido
Estrutura regular	Estrutura irregular
Estrutura independente dos dados	Estrutura embutida no dado
Estrutura reduzida	Estrutura extensa
Estrutura fracamente evolutiva	Estrutura fortemente evolutiva
Estrutura prescritiva	Estrutura descritiva
Distinção entre estrutura e dado é clara	Distinção entre estrutura e dado não é clara

Fonte: Elaborado por MELLO et al. (s.d., p.3)

como, por exemplo, podem ser armazenados em formato de *Extensible Markup Language* (XML) e, através de uma linguagem própria de consulta XML como a XQuery, é possível recuperar os dados armazenados.

XML é uma linguagem de marcação que também é útil na modelagem de dados semi-estruturados. Ela permite, por meio de suas *tags*, definir a estrutura dos dados e a forma como eles se organizam, definindo-se os elementos presentes no documento XML e o conteúdo dos mesmos. Desta forma, é possível adicionar uma certa estrutura a documentos semi-estruturados (PINTO; SACCOL, 2003).

Um exemplo de XML pode ser visto na Figura 2.5 que exemplifica a definição de um arquivo XML para armazenar as informações de um livro, como nome da editora, nome do autor e título do livro.

Figura 2.5 – Exemplo de arquivo XML

```
<book>
      <publisher>
            <name>Morgan Kaufman</name>
      </publisher>
      <author> Author Name</author>
      <title>Title of Book</title>
</book>
```

Fonte: elaborado por WIECZOREK (s.d., p. 2)

Um fator importante ao se utilizar XML para modelar dados semi-estruturados e armazená-los em banco de dados é apresentado por PINTO e SACCOL (2003, p. 2):

> "A definição de um esquema para banco de dados que armazenam documentos semi-estruturados está relacionado à definição de esquema do próprio documento XML. Portanto, o entendimento e a escolha da melhor forma de definir a estrutura de documentos XML pode influenciar na forma com que esses documentos serão armazenados e posteriormente manipulados pelo sistema gerenciador de

> banco de dados ou ainda interferir no processo de troca de informações entre sistemas diferentes. A escolha de uma estrutura que defina satisfatoriamente a estrutura de um documento XML envolve bem mais do que os recursos que serão disponibilizados por tal escolha. A finalidade da estrutura de definição de documentos XML é um fator relevante para sua escolha e que deve ser considerado."

Este presente trabalho utiliza dados semi-estruturados para representar os dados extraídos de notícias, coletadas a partir de páginas *Web*, contendo informações sobre violência contra pessoas LGBTQIA+; em seguida armazena-os em um arquivo XML, para posterior consulta e manipulação no intuito de gerar dados estatísticos sobre violência.

2.1.6 Visualização de Dados

A visão é um dos meios utilizados pelo ser humano para compreensão do mundo ao seu redor. É por meio dela, por exemplo, que consegue fazer distinção de cores, distâncias, formas e movimentos. Além disto, a visualização torna possível a identificação de informações visuais em diversos contextos, tais como: interpretação de linguagem corporal e expressões faciais, orientação através de mapas e compreensão de informações presentes em textos. Além de ser um meio de obtenção de informações, a visualização também possibilita a transmissão das mesmas e, segundo MACEDO et al. (2020 apud BERINATO, s.d., p. 2), pode ser "utilizada como um mecanismo para representar dados sobre o mundo, de forma a diminuir a complexidade do processo de transmissão da informação.".

Quando se visa o contexto da computação, dados podem ser definidos como um meio de estruturar, organizar, identificar e representar informações presentes no mundo. Dados são úteis nos mais diversos e variados contextos como, por exemplo, na realização de transações bancárias, troca de mensagens entre pessoas por meio de dispositivos móveis, coleta de informações por sensores em fábricas, entre outros. Estes dados costumam estar representados em padrões predefinidos, muitas vezes abstratos, necessários para uma comunicação efetiva entre sistemas de software. É necessário que, através de algum meio, o ser humano consiga visualizar e compreender as informações que estão sendo representadas nesses dados. Segundo MACEDO et al. (2020 apud NASCIMENTO; FERREIRA, 2005, p. 2) visualização da informação é "a ciência que estuda como exibir dados abstratos visualmente de maneira que tendências, padrões e relações entre estes dados sejam compreendidas e que *insights* possam ser gerados a partir deles.".

Segundo LOUSA, PEDROSA e BERNARDINO (2019), muitos dos padrões abstratos de dados são difíceis de visualizar e de serem compreendidos, devido à variedade de estruturas de dados e a complexidade dos mesmos; outro fator é o funcionamento do cérebro humano, que possui dificuldade, por exemplo, em encontrar padrões e obter facilmente resultados de relatórios e tabelas de dados. Como forma de facilitar esse processo de visualização, os autores citam a

existência de ferramentas de *Business Intelligence* (BI) para visualização de dados[6] que têm, como objetivo, "reduzir o tempo de interpretação dos dados e permitir a percepção de padrões e relações de dados relevantes"(LOUSA; PEDROSA; BERNARDINO, 2019 apud STODDER, 2015, p. 1). Estas ferramentas permitem, em suas várias formas de visualização, a representação de dados quantitativos em gráficos e tabelas de forma interativa.

Um exemplo de ferramenta de BI para visualização de dados é a *Qlik Cloud*, uma plataforma na nuvem, que permite analisar, visualizar e compartilhar dados, que também utiliza recursos de Inteligência Artificial (IA) fornecendo suporte à tomada de decisões através da associação de dados (CAMPOS BRENO A.; CAMPOS, s.d.). A plataforma aceita diversos tipos de entrada de dados como, por exemplo, arquivos .xml, .csv, .txt, planilhas e bancos de dados, e tem como saída gráficos e tabelas de variados tipos para visualização. Para ter-se uma visualização mais proveitosa dos dados, é necessário organizar e estruturar previamente os mesmos de forma adequada. Segundo LOUSA, PEDROSA e BERNARDINO (2019, p. 2), este processo é conhecido como *Data Preparation* que consistem na "preparação de dados através da combinação, estruturação e organização de dados com objetivo de serem melhor trabalhados.".

A ferramenta *Qlik Cloud* é utilizada neste presente trabalho para apresentação, por meio de gráficos interativos, de dados estatísticos sobre violência contra a população LGBTQIA+ brasileira que foram previamente obtidos por meio de processos de coleta, extração, integração e estruturação XML. A escolha desta ferramenta foi influenciada por proporcionar funcionalidades que atendem os requisitos para alcançar o objetivo do presente trabalho.

2.2 Trabalhos Relacionados

Não foram encontrados na literatura trabalhos que tratam, diretamente, a geração de dados estatísticos sobre violência contra a população LGBTQIA+ brasileira. Porém, foram encontrados trabalhos com contextos e processos semelhantes aos do presente trabalho; alguns encontram-se descritos a seguir.

MEDINA (2022) propôs o desenvolvimento de uma plataforma digital chamada Femini para o fornecimento de dados, informações e visualizações sobre Violência contra Mulher, nos contextos de São Luís, Maranhão e Brasil. Para realizar a coleta e extração dos dados para a construção das visualizações, definiu primeiramente os locais de buscas. Para isso, aplicou um questionário com jornalistas onde uma das perguntas tratava sobre as fontes que os profissionais utilizavam para a produção de notícias sobre Violência contra Mulher. A partir da análise das respostas, em conjunto com buscas on-line no Google por órgãos institucionais que trabalham no combate à Violência contra Mulher, utilizando de operadores de busca avançada, definiu quais instituições seriam contactadas para a coleta de dados. Por meio de notícias que encontrou

[6] De acordo com LOUSA, PEDROSA e BERNARDINO (2019) " ferramentas de BI de visualização de dados são um conjunto de mecanismos que fornecem informação essencial para a tomada de decisão suportados por gráficos, tabelas, *dashboards*[...]".

depois de definir palavras-chave para sua busca, procurou identificar manualmente nas notícias quais eram as instituições que compunham as fontes de dados das reportagens. Nos sites das instituições, coletou alguns dados através de formatos PDF e CSV, e tentou coletar dados através de contato direto com algumas instituições municipais e estaduais. A autora percebeu uma não sistematização de dados durante a fase da coleta, onde cada órgão era responsável por organizar os dados de acordo com seus regimentos internos, não havendo um padrão geral para dados de Violência contra Mulher.

Para construir a visualização dos dados, MEDINA (2022) criou uma página web e, por meio de ferramentas como *Piktochat*, criou infográficos [7] e mapas, no intuito de apresentar os dados de maneira mais dinâmica e interativa, a partir dos dados coletados. Para produzir os mapas, utilizou a ferramenta *My Maps* do *Google* e necessitou adicionar manualmente os números de casos em cada estado e utilizou de um processo de adição automatizada para os casos que ocorreram em cada bairro de São Luís. Todos esses casos foram representados através de marcadores nos mapas que a ferramenta escolhida já oferece. MEDINA (2022) também disponibilizou na Femini uma opção para fazer *download* de uma das bases de dados utilizadas e uma lista de *link* para redirecionamentos para outras bases, com o objetivo de que os usuários pudessem acessar os dados de maneira mais rápida e em um só lugar. Outra funcionalidade disponibilizada na plataforma foi um formulário para inscrição em uma *newsletter*, com a intenção de proporcionar aos usuários novidades e atualizações sobre as visualizações, os dados presentes na plataforma e outras informações sobre Violência Contra Mulher.

As dificuldades encontradas no trabalho de MEDINA (2022) envolveram o acesso e extração dos dados. No caso do acesso aos dados, pouquíssimas instituições possuíam uma base de dados e a disponibilizavam no meio digital, e quando o faziam, muitas eram no formato PDF que dificultava a coleta e análise dos mesmos, tornando uma tarefa longa e manual quando esta poderia ser automatizada. A extração tornou-se complexa devido ao fato de cada instituição registrar os dados a sua maneira, resultando em tabelas com registros diferentes de uma mesma localidade. A visualização dos dados foi reportada como um ponto de cuidado pela autora: a mesma alegou que representações visuais podem levar a interpretações errôneas do que os números obtidos realmente representam e que uma análise mal feita dos dados pode resultar em péssimos gráficos, mapas imprecisos ou infográficos que transmitem algo distante do que era intencionado.

WANG et al. (2021) propuseram uma abordagem para extração de informações de notícias online chinesas por meio de um sistema chamado *Chinese News Fact Extractor* (CNFE). Para os autores, existe um conceito importante na coleta de informações de notícias, onde originalmente uma notícia deve ser considerada completa se responder a uma lista de seis perguntas: o quê, por quê, quem, quando, onde e como. Com apoio destas perguntas, propuseram um método de

[7] De acordo com TEIXEIRA (2010) infográfico "é composto por elementos icônicos e tipográficos e pode ser constituído por mapas, fotografias, ilustrações, gráficos e outros recursos visuais, inclusive aqueles mais abstratos e não necessariamente icônicos".

extração semântica de eventos [8] de notícias. Foram duas as principais contribuições deste trabalho: a primeira foi a proposta de um algoritmo para extrair frases temáticas a partir da estrutura de uma notícia, principalmente de sua manchete; a segunda foi a extração de fatos de eventos de notícias onde, neste contexto, fatos são as respostas para as seis perguntas propostas, por meio de um método de aprendizado supervisionado[9]. Este método melhorou significativamente na Extração Automática de Conteúdo (EAC) de notícias escritas em chinês e apresentou uma escalabilidade muito alta devido ao fato de considerar apenas as sentenças dos tópicos e as caraterísticas da superfície do texto. Com base neste método, implementaram o CNFE e avaliaram a precisão de sua extração automática de conteúdo com 30.000 notícias reais, onde os resultados dos experimentos mostraram que as informações puderam ser extraídas com eficácia.

FORNARI et al. (2021) propuseram uma análise de como as mídias digitais retratam a violência contra a mulher no início da pandemia da COVID-19, no Brasil, à luz de gênero. Para isso, realizaram um estudo descritivo de abordagem qualitativa utilizando dados *online* (notícias e comentários) publicados em plataformas digitais, estas sendo: portais de notícias, jornais, sites governamentais e de organizações feministas e a rede social *Twitter*. A escolha das plataformas deu-se pela possibilidade de captar diferentes perspectivas do objeto estudado. O acesso às notícias e aos comentários deu-se por meio do campo de busca de cada site investigado, usando termos específicos e excluindo os resultados que apresentaram duplicidade. Salvaram as notícias e comentários selecionados no formato PDF, armazenados e compartilhados entre as autoras por meio de pastas de arquivos *online*. Posteriormente, fizeram a extração dos dados por meio de instrumento semiestruturado, desenvolvido pelas autoras para uso exclusivo do estudo, com o objetivo de capturar informações específicas, sendo estas: data de publicação e autoria da notícia, plataforma digital e discursos sobre violência contra a mulher. Logo após, trataram-os pela análise de conteúdo temática, com suporte do *software* webQDA. A análise de conteúdo constitui-se das etapas: a) pré-análise, definida pela organização e leitura do material; b) exploração do material, constituída pela codificação e categorização; c) tratamento dos resultados. Os dados, depois de coletados pelo instrumento, foram inseridos em uma planilha Excel e incorporados ao *software* de análise qualitativa *webQDA*, que foi utilizado por permitir o tratamento dos dados de maneira colaborativa e síncrona entre a equipe de pesquisa. Por meio da análise manual dos dados obtidos, especificamente a leitura, encontraram três categorias empíricas: os reflexos da COVID-19 nos números da violência contra a mulher; a COVID-19 desvelando a violência contra a mulher no público e no privado; COVID-19 e violência contra a mulher: duas pandemias em paralelo. A análise realizada evidenciou a potencialidade do meio digital para compreender a expressão da violência contra a mulher durante a pandemia do COVID-19, estimulando reflexões que podem

[8] Para WANG et al. (2021) evento é o ponto central da notícia, ou seja, o tema que esta está retratando.

[9] Segundo MONARD e BARANAUSKAS (2003) *aprendizado supervisionado* consiste em fornecer ao algoritmo de aprendizado ”um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido” e o algoritmo de aprendizado supervisionado tem como objetivo ”construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham o rótulo da classe”.

vir a contribuir para o seu enfrentamento.

Diferente do trabalho atual, nenhum dos trabalhos citados realizou coleta semiautomática das notícias relevantes a seu contexto de pesquisa. Semelhante a este trabalho, WANG et al. (2021) utilizaram da extração automatizada de informações a partir de notícias, FORNARI et al. (2021) utilizaram de dados semiestruturados para modelagem dos dados extraídos, e MEDINA (2022) permitiu o compartilhamento e a apresentação dos dados obtidos, mesmo que de forma não automatizada. Neste presente trabalho, diferentemente dos citados, todas as etapas de coleta, extração, integração e apresentação de dados serão automatizadas. A coleta possui como temática específica a violência contra a população LGBTQIA+ brasileira; ademais, os dados a serem obtidos referem-se as estatísticas de tais violências, sendo apresentados ao final de forma automatizada, dinâmica e interativa.

3 Desenvolvimento

Como apresentado na Seção 1.2, este presente trabalho tem, como objetivo geral, propor e desenvolver uma primeira versão completa e funcional de uma estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira. Para tal, têm-se, como base: os conceitos definidos nas Subseções 2.1.3, 2.1.4, 2.1.5 e 2.1.6; a Seção 3.1 que descreve em detalhes a arquitetura de funcionamento da estratégia; e, por fim, a Seção 3.2 que expõe e exemplifica as interfaces da ferramenta inicial que implementa a estratégia proposta.

3.1 Arquitetura de funcionamento da estratégia proposta

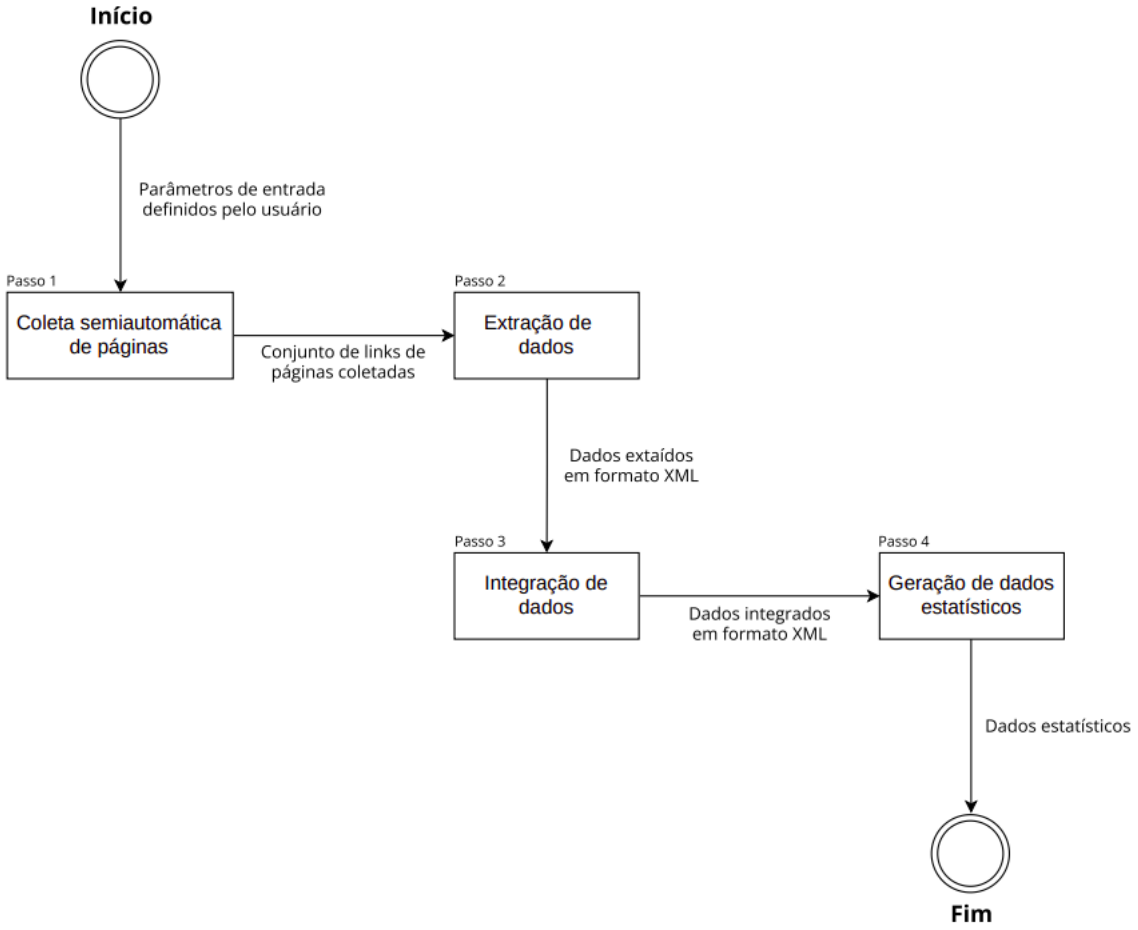
A arquitetura de funcionamento da estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira encontra-se definida na Figura 3.1.

Como pode ser observado na Figura 3.1, a arquitetura de funcionamento apresenta 4 passos e eles definem-se, de forma geral, como:

- **Passo 1 - Coleta semiautomática de páginas:** consiste em, por meio de consultas realizadas programaticamente no serviço de busca da *Google*, coletar páginas *Web* que contenham notícias sobre casos de violência contra pertencentes da comunidade LGBTQIA+ brasileira. Para isso, têm-se, como parâmetros de entrada definidos pelo usuário: um arquivo contendo os perfis LGBTQIA+ que se deseja encontrar nas notícias; um arquivo contendo o conjunto de termos caracterizando violências a serem associados aos perfis na busca; um conjunto de domínio dos sites de notícia onde as notícias podem ser buscadas; e a quantidade de notícias a serem buscadas. O algoritmo de coleta realiza então combinações entre os dados fornecidos pelo usuário e, por meio destas combinações, pesquisa no buscador da *Google* e retorna para o usuário uma lista contendo os *links* encontrados. O usuário consegue então avaliar e selecionar os *links* que realmente dizem respeitos a casos de violência contra LGBTQIA+ brasileiros. O conjunto de *links* selecionados é então armazenado em um arquivo de texto que representa a entrada do Passo 2.

- **Passo 2 - Extração de dados:** consiste em extrair, a partir de cada página *Web* do conjunto de páginas retornado pelo Passo 1, dados relevantes sobre violência contra pessoas LGBTQIA+ brasileiras. Por meio de técnicas de extração por abordagem manual (vide Subseção 2.1.4) combinadas com mineração de texto, os dados são extraídos da página e modelados de forma semi-estruturada em formato XML. Os dados em formato XML

Figura 3.1 – Arquitetura de funcionamento da estratégia proposta para geração de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira



Fonte: elaborado pela autora

geram um arquivo .XML para cada notícia contendo as informações extraídas de cada página *Web*. Estes dados são então enviados como entrada para o Passo 3.

- **Passo 3 - Integração de dados:** consiste em integrar os dados XML obtidos e armazená-los em um único arquivo XML que constitui um banco de dados semi-estruturado. Uma inteligência é aplicada para analisar em conjunto os arquivos obtidos como entrada, contendo os dados extraídos em formato XML, e comparar com os dados já armazenados, com o propósito de não armazenar informações repetidas e complementar, quando necessário, as já existentes no banco de dados. O banco de dados semi-estruturado, um arquivo .XML contendo todos os dados de todas as notícias em formato XML, é gerado durante o processo de integração dos dados de todos os arquivos retornados pelo Passo 2 e serve então como entrada para o Passo 4.

- **Passo 4 - Geração de dados estatísticos:** consiste em gerar dinamicamente dados estatísticos a partir de consultas e associações sobre os dados presentes no banco de dados semi-estruturado gerado no Passo 3. Utilizando-se da ferramenta de BI *Qlik Cloud* (vide

Subseção 2.1.6), o arquivo XML, que consiste no banco de dados semi-estruturado contendo informações de casos de violência contra LGBTQIA+ brasileiros, é fornecido como entrada e a ferramenta realiza associações de dados para geração de dados estatísticos. Na ferramenta, é possível indicar os dados a serem associados, gerar gráficos e filtros interativos para a visualização dos dados. Os dados estatísticos obtidos representam o objetivo final da presente estratégia e poderão ser visualizados e compartilhados.

Os componentes dos Passo 1, Passo 2 e Passo 3, presentes na arquitetura de funcionamento apresentada na Figura 3.1, merecem uma melhor descrição e, desta forma, as próximas subseções desta seção detalham o funcionamento de cada um destes componentes. A Subseção 3.1.1 detalha como é feita a coleta semiautomática de páginas da presente estratégia e refere-se ao Passo 1 da arquitetura de funcionamento. Já a Subseção 3.1.2 detalha como é feito o processo de extração de dados das páginas de notícia de violências contra a comunidade LGBTQIA+ brasileira. Esta subseção refere-se ao Passo 2 da arquitetura de funcionamento. Por fim, a Subseção 3.1.3 explora em detalhes como é feita a integração dos dados obtidos pelo processo de extração e refere-se ao Passo 3 da arquitetura de funcionamento.

3.1.1 Coleta semiautomática de páginas

Nesta seção, é descrito, em detalhes, como foi realizada a coleta de páginas *Web* de notícias sobre casos de violência contra pertencentes da comunidade LGBTQIA+ brasileira, representando o componente do Passo 1 da arquitetura de funcionamento (vide Figura 3.1).

Para a realização da coleta de páginas de notícias, foi implementado um algoritmo em Python responsável por realizar pesquisas programaticamente no buscador da *Google* utilizando a biblioteca *googlesearch*. A biblioteca *googlesearch* é uma API [1] *Python* de código aberto que permite realizar pesquisas no *Google* de forma programática, facilitando a coleta automatizada de informações do mecanismo de busca para diversos fins. A utilização desta biblioteca possui algumas restrições quanto a limite de solicitações permitidas por dia ou por unidade de tempo.

Antes da execução do algoritmo de coleta, o usuário deve definir os parâmetros de entrada do algoritmo:

- Especificar em um arquivo texto os perfis de indivíduos LGBTQIA+ (por exemplo, lésbicas, gays e travestis) que devem ser encontrados nas notícias de casos de violência. O usuário deve especificar em cada linha do arquivo um perfil que deseja que esteja presente nos resultados da coleta.

[1] De acordo com Pereira e Cabral (2023), uma API é ”uma sigla para *Application Programming Interface* (interface de programação de aplicação). No contexto de APIs, a palavra aplicação refere-se a qualquer *software* com uma função distinta. A interface pode ser pensada como um contrato de serviço entre duas aplicações. Esse contrato define como as duas se comunicam usando solicitações e respostas. Uma API inclui mecanismos que permitem que dois componentes de *software* comuniquem-se usando um conjunto de definições e protocolos.”.

- Especificar em um arquivo texto os termos de violência que deseja associar a cada perfil LGBTQIA+ para realização da busca de notícias. São termos que muito provavelmente estarão presentes em notícias de violência como, por exemplo, assassinado, espancado, agredido, etc. O usuário deve especificar em cada linha do arquivo um termo de violência distinto.

- Especificar em um arquivo texto os domínios de sites de notícias em que deseja encontrar as notícias. A ferramenta de busca da *Google* permite que buscas sejam realizadas em sites específicos, desde que o domínio do site seja especificado na busca através de uma estrutura específica como, por exemplo, "site: g1.com.br". O usuário deve especificar em cada linha do arquivo um domínio de um site distinto.

O algoritmo de coleta semiautomática de páginas possui uma interface de interação com o usuário e consiste nos seguintes passos:

- Passo 01: escolher o arquivo contendo os perfis LGBTQIA+ a serem encontrados em notícias de casos de violência (tarefa do usuário);

- Passo 02: selecionar os perfis LGBTQIA+ a serem buscados em notícias de casos de violência (tarefa do usuário). Após o usuário carregar o arquivo com os perfis no Passo 01, ele pode selecionar quais dos perfis estarão presentes na busca, sem necessariamente realizar uma pesquisa com todos os perfis presentes no arquivo;

- Passo 03: escolher o arquivo contendo os termos de violência que deseja associar a cada perfil LGBTQIA+ para realizar a busca por notícias de violência (tarefa do usuário);

- Passo 04: selecionar os termos de violência que estarão associados a cada perfil LGBTQIA+ selecionado no Passo 02 (tarefa do usuário). Após o usuário carregar o arquivo com os termos de violência no Passo 03, ele pode selecionar quais dos termos estarão presentes na busca, sem necessariamente realizar uma pesquisa com todos os termos presentes no arquivo;

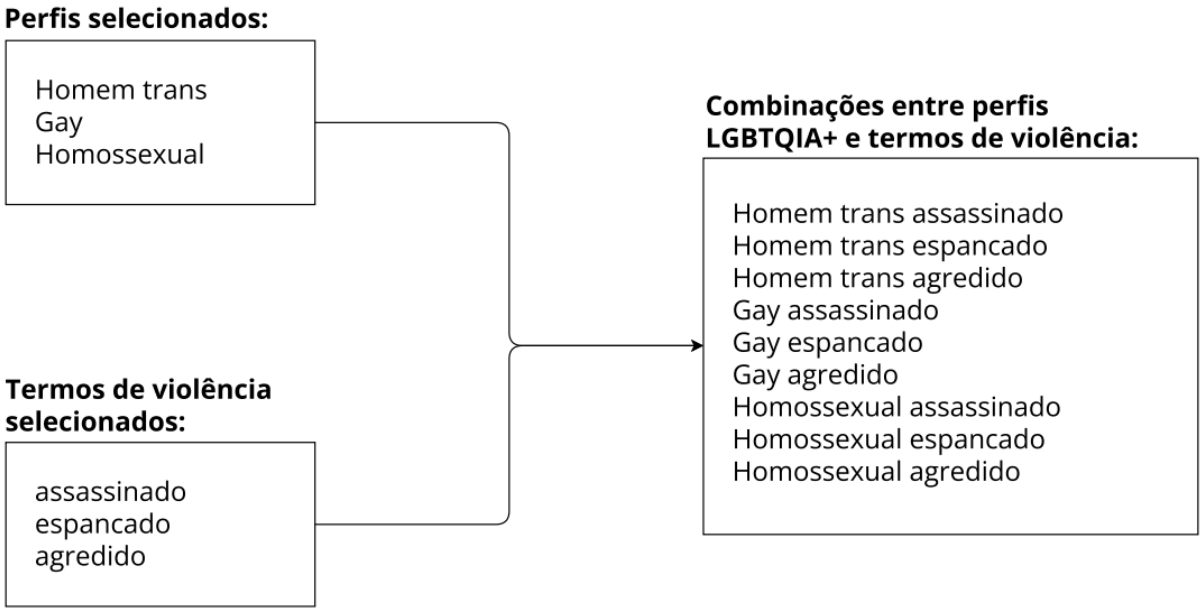
- Passo 05: escolher o arquivo contendo os domínios dos sites em que se deseja encontrar as notícias de violência (tarefa do usuário);

- Passo 06: selecionar os domínios dos sites em que se deseja encontrar as notícias de violência (tarefa do usuário). Após o usuário carregar o arquivo com o domínio dos sites no Passo 05, ele ainda pode optar por escolher entre os domínios presentes no arquivo, sem necessariamente tentar realizar pesquisas por notícias em todos eles;

- Passo 07: especificar a quantidade de notícias que devem ser retornadas pelo algoritmo coletor (tarefa do usuário);

- Passo 08: combinar as informações de perfil LGBTQIA+ e termos de violência carregadas e selecionadas pelo usuário para realizar a coleta de notícias de violência. O algoritmo primeiramente identifica todos os termos de perfis LGBTQIA+ selecionados pelo usuário no Passo 02. Em seguida, combina cada perfil LGBTQIA+ com cada termo de violência selecionado no Passo 04. O total de combinações de perfis e termos de violência é definido pela quantidade de perfis selecionados multiplicada pela quantidade de termos de violência selecionados. Um exemplo das combinações produzidas no Passo 08 pode ser visualizado na Figura 3.2;

Figura 3.2 – Combinações entre perfis LGBTQIA+ e termos de violência



Fonte: elaborado pela autora

- Passo 09: definir quantos *links* o coletor irá obter de cada pesquisa realizada por cada combinação entre perfil LGBTQIA+ e termo de violência no buscador da *Google*. Ao utilizar a biblioteca *googlesearch*, é possível realizar uma consulta no buscador por meio do método *search* que recebe como parâmetro o conjunto de termos a serem buscados e a quantidade de *links* que serão retornados na consulta. A quantidade de *links* retornados a cada pesquisa por cada combinação gerada no Passo 09 é definida pelo algoritmo coletor. Primeiramente, verifica-se se a quantidade de notícias especificadas pelo usuário no Passo 07 é menor que a quantidade de combinações entre perfis LGBTQIA+ e termos de violência gerada no Passo 09. Caso seja este o caso, o algoritmo seleciona aleatoriamente combinações até que a quantidade selecionada satisfaça o número de notícias especificados pelo usuário e define que, para cada termo selecionado, será buscado exatamente um único *link* correspondente a uma notícia de violência. Se a quantidade de notícias especificadas pelo usuário for exatamente igual a quantidade de combinações geradas entre perfil LGBTQIA+ e termos violências, o algoritmo define que para cada combinação de termos um único *link* correspondente a uma notícia será retornado. Se os dois casos anteriores forem falsos,

indica que a quantidade de notícias especificadas pelo usuário é maior que a quantidade de combinações geradas entre perfil LGBTQIA+ e termos violências. Neste caso, o algoritmo divide a quantidade de notícias especificadas pela quantidade de termos e define que o resultado da divisão será a quantidade de páginas retornadas para cada combinação de termos. Se o resultado da divisão não for exato, ou seja, o restante da divisão não resultar em zero, o algoritmo escolhe aleatoriamente uma das combinações de termos e adiciona o restante da divisão em seu valor que indica a quantidade de *links* a serem retornados. O algoritmo sempre garante que a quantidade de notícias solicitadas pelo usuário será retornada e realiza as técnicas descritas visando garantir que exista um equilíbrio nas buscas entre os perfis LGBTQIA+, termos de violência e quantidade de notícia informados pelo usuário. Este equilíbrio visa fazer uma média de quantos *links* serão retornados pela busca de cada combinação entre perfil LGBTQIA+ e termo de violência, de modo que a quantidade retornada por combinação não fique muito discrepante em relação as outras e atenda os requisitos de busca informados pelo usuário;

- Passo 10: realizar a pesquisa por notícias de violência no buscador da *Google* através das combinações definidas no Passo 08 e da quantidade de *links* a serem retornados por combinação definida no Passo 09. Para cada combinação, o algoritmo seleciona aleatoriamente um dos domínios de sites de notícias selecionados pelo usuário no Passo 06. Esta seleção aleatória de um domínio entre os selecionados pelo usuário visa uma maior diversidade de notícias, já que cada busca por combinação de termo poderá ser feita em diferentes domínios. Após selecionar o domínio aleatoriamente, o algoritmo realiza então a combinação de termos que representará o termo a ser buscado fornecido ao método *search* da biblioteca *googlesearch*. Cada termo de busca, fornecido como parâmetro ao método, constitui-se da seguinte forma: "site: <domínio escolhido aleatoriamente> <combinação de perfil LGBTQIA+ e termo de violência>". O termo é fornecido ao método *search* em conjunto com a quantidade de *links* (definida no Passo 09) que deve ser retornada na pesquisa. Para cada combinação gerada no Passo 08, em conjunto com a quantidade de *links* a ser retornada obtida no Passo 09, é realizada uma pesquisa no Google e o seu resultado armazenado. Enquanto as buscas estão sendo realizadas, um componente de *loading* informa ao usuário que elas estão ocorrendo. Depois de contemplar a busca por todas as combinações, o algoritmo combina os resultados em um conjunto de *links* que contém exatamente a quantidade de *links* solicitadas pelo usuário no Passo 07 e os apresenta na interface para o usuário;

- Passo 11: analisar e selecionar os *links* que considerar serem referentes a notícias de violências contra LGBTQIA+ brasileiros (tarefa do usuário). Como descrito anteriormente, o Passo 10 é responsável por realizar a busca por notícias de casos de violência contra LGBTQIA+ e exibir os resultados encontrados ao usuário. Neste momento, o usuário analisa cada *link* retornado e seleciona os *links* que considera adequados;

- Passo 12: armazenar os *links* selecionados pelo usuário no Passo 11. Após selecionar os *links* no Passo 11, o algoritmo de coleta armazena então os *links*, adicionando cada *link* em uma linha ao final do arquivo que contém o conjunto de *links*, referentes a notícias de casos de violência contra a comunidade LGBTQIA+ brasileira, coletados pelo algoritmo;
- Passo 13: remover os *links* duplicados que possam existir. Como pode acontecer de diferentes execuções do algoritmo de coleta trazerem como resultado *links* repetidos e o usuário acabar selecionando *links* já escolhidos em outras execuções do algoritmo, optou-se por tratá-los logo após serem armazenados no arquivo, removendo toda repetição de um *link* já existente.

Ao final dos 13 passos descritos, tem-se então um arquivo contendo *links* de páginas de notícias de violência contra LGBTQIA+ brasileiros. Este arquivo compõe, então, o conjunto de *links* de páginas coletadas que serve como entrada para o Passo 2 da arquitetura de funcionamento (vide Figura 3.1), como já mencionado.

3.1.2 Extração de dados

A presente seção descreve, em detalhes, como é feita a extração de dados relevantes de páginas *Web* referentes a notícias sobre casos de violência contra LGBTQIA+ brasileiros, representando o componente do Passo 2 da arquitetura de funcionamento, representada na Figura 3.1. A partir do conjunto de páginas coletadas no Passo 1 (vide Figura 3.1), são extraídos os dados relevantes referentes a violência contra a população LGBTQIA+ de cada uma das páginas presentes no conjunto, por meio da extração por abordagem manual (vide Subseção 2.1.4).

O primeiro passo da extração de dados das páginas *Web* de notícias de violência contra LGBTQIA+ brasileiros, é a obtenção dos textos que descrevem os fatos noticiados. Geralmente, as páginas de notícias possuem em sua estrutura: um título (também conhecido como manchete); o corpo da notícia, onde detalhes e informações sobre o fato ocorrido são apresentados; fotos; vídeos; data e hora da publicação ou atualização; e, também, anúncios. Este trabalho considera que dados referentes a casos de violência contra LGBTQIA+ brasileiros estão localizados na estrutura textual da notícia, sendo assim essencial obter essa estrutura para a realização do processo de extração.

Para a obtenção do texto das páginas e a realização de todo o processo de extração, optou-se pela utilização da linguagem de programação *Python*, devido à disponibilidade de recursos que a mesma oferece. Utilizando-se da biblioteca *BeautifulSoup* disponível na linguagem, biblioteca esta que permite a análise e extração de dados em documentos HTML, extrai-se toda a estrutura textual das páginas de notícia. Para cada *link*, presente no conjunto de *links* retornado pelo Passo 1 da arquitetura arquitetura de funcionamento (vide Figura 3.1), uma requisição é feita utilizando a biblioteca descrita, com o intuito de obter o código HTML que estrutura a página *Web* da notícia. Caso a requisição falhe por algum motivo, por exemplo, o site não permitir que a informação

textual seja coletada, o próximo *link* do repositório é visitado. Caso a requisição tenha sucesso e o código HTML seja retornado, a biblioteca permite então que o conteúdo HTML seja refinado, possibilitando a exclusão de algumas *tags* HTML, seja pelo tipo da *tag* ou por classes associadas a ela, e também de textos indesejados.

É notável, ao se avaliar notícias sobre qualquer assunto, que as mesmas não possuem um padrão exato do conteúdo que devem apresentar. Um mesmo fato noticiado em diferentes fontes pode apresentar informações distintas: uma fonte pode possuir informações mais completas que outras. Tendo esta percepção, definiu-se então quais seriam os dados mais relevantes a serem coletados de notícias de casos de violência contra a população LGBTQIA+, considerando-se que todos podem estar presentes ou não no texto obtido das notícias coletadas. Estes dados são:

- **Data em que a notícia foi publicada:** dado necessário para levantamento de informações estatísticas sobre o índice de casos de violência contra LGBTQIA+ brasileiros em relação a meses e anos. Também auxilia no processo de integração dos dados;
- **Nome da vítima:** dado necessário para extração de outras informações referentes a vítima e no processo de integração dos dados (vide Subseção 3.1.3), onde auxilia a evitar o armazenamento de dados duplicados;
- **Idade da vítima:** dado necessário para o levantamento de informações estatísticas sobre a faixa etária das vítimas de violência e auxilia no processo de integração dos dados;
- **Orientação sexual da vítima:** dado necessário para o levantamento de informações estatísticas sobre a orientação sexual das vítimas de violência e auxilia no processo de integração dos dados;
- **Identidade de gênero da vítima:** dado necessário para levantamento de informações estatísticas sobre a identidade de gênero das vítimas de violência e auxilia no processo de integração dos dados;
- **Bairro em que a vítima sofreu violência:** dado necessário para levantamento de informações estatísticas sobre índices de violência contra LGBTQIA+ em regiões específicas de cidades;
- **Cidade em que a vítima sofreu violência:** dado necessário para levantamento de informações estatísticas sobre índices de violência contra LGBTQIA+ em cidades brasileiras;
- **Estado em que a vítima sofreu violência:** dado necessário para levantamento de informações estatísticas sobre índices de violência contra LGBTQIA+ em estados brasileiros e auxilia no processo de integração dos dados;
- **Região do Brasil em que a vítima sofreu violência:** dado necessário para levantamento de informações estatísticas sobre índices de violência contra LGBTQIA+ nas regiões do Brasil e auxilia no processo de integração dos dados;

- **Tipos de violência sofridos pela vítima:** dado necessário para levantamento de informações estatísticas sobre os tipos de violências mais frequentes contra LGBTQIA+ brasileiros.

Tendo como objetivo a extração dos dados citados, definiu-se então as estratégias para a extração de cada um deles. O algoritmo de extração de dados foi implementado na linguagem *Python* e existem diferentes técnicas de extração para cada um dos dados descritos anteriormente. O algoritmo em uma única execução retorna, quando encontrados e extraídos a partir das técnicas implementadas, os dados descritos. Para cada *link* do conjunto de *links* de entrada, primeiramente, o algoritmo obtém o texto da notícia utilizando a biblioteca *BeautifulSoup*, conforme descrito anteriormente. A partir do texto da notícia obtido, todo processo de extração de dados é então iniciado. Após a extração dos dados citados, quando estes existem, um arquivo XML, como o que pode ser visualizado na Figura 3.3, é construído para cada notícia, contendo o *link* da notícia a partir dos quais os dados foram extraídos e os dados extraídos. A biblioteca *xml.etree.ElementTree* da linguagem *Python* é utilizada para estruturar os dados extraídos em formato XML. Cada dado recebe uma *tag* que identifica a qual tipo o dado armazenado pertence, por exemplo, a idade da vítima é armazenada através da *tag* <idade>. Ao final do processo de extração de dados, têm-se um conjunto de arquivos .XML, contendo todas as informações extraídas de todas as notícias, que representa o conjunto de dados que serve como entrada para o Passo 3 da arquitetura de funcionamento (vide Figura 3.1).

Figura 3.3 – XML resultante do processo completo de extração a partir de uma notícia coletada

```
▼<root>
 ▼<noticia>
    <link>https://ultimosegundo.ig.com.br/brasil/2020-08-19/vou-levar-essas-marcas-
    eternamente-diz-modelo-trans-agredida-em-copacabana.html</link>
    <id>53</id>
    <nome>Alice Felis</nome>
    <identidade_de_genero>trans</identidade_de_genero>
    <cidade>Rio de Janeiro</cidade>
    <estado>Rio de Janeiro</estado>
    <regiao>Sudeste</regiao>
    <dia>19</dia>
    <mes>8</mes>
    <ano>2020</ano>
   ▼<violencias>
    ▼<agressao_fisica>
       <termos>soco</termos>
       <termos>agredid</termos>
       <termos>agressão</termos>
     </agressao_fisica>
    ▼<agressao_verbal>
       <termos>xingad</termos>
     </agressao_verbal>
    </violencias>
  </noticia>
</root>
```

Fonte: elaborado pela autora

Visando explanar de forma mais clara como a extração de cada dado ocorre, cada processo e técnica de extração estão detalhados nas subseções a seguir. A Subseção 3.1.2.1 descreve sobre

o processo de extração da data em que a notícia foi publicada. A Subseção 3.1.2.2 aborda o processo de extração do nome da vítima. Por sua vez, a Subseção 3.1.2.3 contempla o processo de extração da idade da vítima, enquanto a Subseção 3.1.2.4 contempla a extração da identidade de gênero e orientação sexual da vítima. Já a Subseção 3.1.2.5 detalha o processo de extração da localidade em que vítima sofreu violência, consistindo por localidade o bairro, a cidade, o estado e a região do Brasil. Por fim, a Subseção 3.1.2.6 descreve como é realizada a extração dos tipos de violência sofridos pela vítima.

3.1.2.1 Extração da data de publicação da notícia

Nesta subseção, encontra-se, detalhadamente, o processo de extração da data que uma notícia foi publicada. A data é um dado importante para compreender os índices de violência contra LGBTQIA+ por meio dos anos e meses. Como uma notícia atualmente é publicada em um espaço de tempo muito próximo a ocorrência do fato que a mesma retrata, optou-se por extrair a data da notícia para identificar a data em que um ato violento ocorreu. Este trabalho considera relevantes apenas o mês e o ano, da data noticiada de uma violência ocorrida contra uma pessoa LGBTQIA+, para o levantamento de dados estatísticos relativos a índices de casos de violência contra LGBTQIA+ brasileiros em relação a meses e anos.

A partir do texto obtido de uma notícia, o algoritmo extrai a data seguindo os seguintes passos:

- Passo 01: extrair a data presente na notícia quando esta está representada apenas em números. A partir dos padrões numéricos de data *dd/mm/aa* ou *dd/mm/aaaa* [2], extrai-se a data utilizando uma expressão regular[3] que identifica este padrão no texto e salva a primeira data encontrada. A partir da análise de textos de notícias, observou-se que a data de uma notícia é uma das primeiras informações presentes e, devido a esta observação, optou-se por escolher a primeira data presente no texto como sendo a data da notícia;
- Passo 02: extrair a data presente na notícia quando esta está representada através de número e texto. Quando o Passo 01 falha e não consegue encontrar uma data no padrão definido, o algoritmo tenta então extrair a data a partir do padrão: dia escrito em dois algarismos numéricos, mês escrito por extenso e ano escrito em dois ou quatro algarismos numéricos. A primeira data encontrada no texto no padrão descrito também é escolhida como a data da notícia, assim como descrito no Passo 01.

[2] Nos padrões descritos, *dd* corresponde ao dia de uma data representado por dois algarismos numéricos como, por exemplo, 01. Já *mm* correspondem ao mês de uma data representado também por dois números. Por fim, *aa* ou também *aaaa*, contemplam os anos de uma data representados por dois ou quatro números respectivamente. Um exemplo de data no padrão *dd/mm/aa* é 20/01/98 e a mesma data no padrão *dd/mm/aaaa* seria representada como 20/01/1998.

[3] Segundo Gomes, D'Emery e Cysneiros (2014), uma expressão regular é "uma maneira de descrever conjuntos regulares. É utilizada em construção de compiladores, editores, sistemas operacionais, protocolos, etc. Trata-se de um formalismo denotacional, também considerado gerador, que permite a inferência sobre a construção ou geração de palavras pertencentes a uma linguagem."

A partir dos dois passos descritos anteriormente, é possível extrair a data do texto de uma notícia de violência contra um indivíduo pertencente a comunidade LGBTQIA+ brasileira.

3.1.2.2 Extração do nome da vítima

Esta subseção descreve, em detalhes, como é feito todo o processo de extração dos nomes de vítimas de violência, pertencentes a comunidade LGBTQIA+ brasileira, a partir de textos de notícias. O nome da vítima trata-se de um dado pessoal e este trabalho não utiliza da extração deste para qualquer forma de identificação ou divulgação de informações pessoais sobre vítimas de violência. A coleta do nome tem como intuito exclusivo auxiliar no processo de integração de dados, visando evitar a duplicidade dos dados armazenados.

Para realização da extração dos nomes das vítimas tem-se, como base, o texto de uma notícia obtido conforme descrito anteriormente. O algoritmo de extração de nomes combina mineração de texto e extração por abordagem manual, sendo constituído pelos seguintes passos:

- Passo 01: utilizar a biblioteca *SpaCy*, da linguagem *Python*, para processar e minerar um texto em português, com o intuito de extrair nomes próprios de pessoas presentes no texto. A *SpaCy* é uma biblioteca de Processamento de Linguagem Natural (PLN) que oferece uma ampla gama de recursos para análise e compreensão de textos em várias línguas e permite a extração de informações específicas, análises gramaticais e semânticas, classificação de textos e outras funcionalidades. O módulo *pt_core_news_lg* e o rótulo "PER"[4] foram utilizados para a extração dos nomes próprios;
- Passo 02: remover os nomes duplicados retornados pelo Passo 01;
- Passo 03: contar a frequência no texto de todos os nomes retornados pelo Passo 02;
- Passo 04: ordenar os nomes pelos nomes mais frequentes identificados pelo Passo 03;
- Passo 05: obter o primeiro nome mais frequente;
- Passo 06: obter o segundo nome mais frequente;
- Passo 07: retornar o nome da vítima. Este passo valida se o segundo nome mais frequente (obtido no Passo 06) é maior que o primeiro (obtido no Passo 05). Caso esta condição não seja verdadeira, o primeiro nome mais frequente é retornado como o nome da vítima; mas, caso a condição seja verdadeira, verifica-se então se o primeiro nome mais frequente está contido no segundo. Caso esteja, o segundo nome mais frequente é retornado como sendo o nome da vítima e, caso não esteja, o primeiro nome mais frequente é retornado.

[4] O rótulo "PER", proveniente do módulo *pt_core_news_lg* da biblioteca *SpaCy*, é uma abreviação da palavra inglesa *"Person"* que significa "Pessoa". Esse rótulo é atribuído por modelos de processamento de linguagem treinados para identificar entidades nomeadas em um texto, e ele indica a presença de um nome de pessoa. Esses modelos utilizam técnicas de aprendizado de máquina para reconhecer automaticamente nomes próprios de indivíduos.

Esta técnica foi implementada por ter-se identificado que, em notícias em que o nome da vítima é citado, geralmente este nome é o que possui mais ocorrências em todo texto. E, caso o segundo nome contenha o primeiro, pressupõe-se que seja o nome da vítima mais completo, contendo seus sobrenomes.

Ao final dos sete passos descritos, têm-se o nome da vítima extraído, caso este tenha sido citado no texto da notícia.

3.1.2.3 Extração da idade da vítima

Nesta subseção, são apresentados os pormenores do procedimento completo da extração das idades das vítimas de violência pertencentes à comunidade LGBTQIA+ do Brasil, a partir de informações contidas em notícias. A extração da idade faz-se necessária para a geração de dados estatísticos sobre a faixa etária das vítimas de violência.

Para a extração da idade das vítimas, têm-se como entrada o texto extraído de uma notícia e o nome da vítima (vide Subseção 3.1.3). O algoritmo de extração de idades constitui-se dos seguintes passos:

- Passo 01: encontrar todas as ocorrências de idades no texto. A partir de uma expressão regular, o algoritmo identifica e armazena todos os números que são sucedidos pela palavra "anos";
- Passo 02: encontrar no texto a idade em um intervalo específico de caracteres em relação ao nome da vítima. Para cada idade encontrada no Passo 01, o algoritmo percorre o texto tentando encontrar, no intervalo de até cem caracteres antes da ocorrência da idade, o nome da vítima. Caso o nome seja encontrado neste intervalo, o algoritmo assume que esta é idade correta da vítima e a retorna, não realizando a busca pelo nome em relação as demais idades encontradas no Passo 01. Esta lógica baseia-se na observação de que, em muitas notícias, quando a idade da vítima é citada e o texto também cita seu nome, a idade ocorre em um intervalo próximo de caracteres logo após a ocorrência do nome. Caso o nome da vítima encontrado não conste nos cem caracteres antecessores de nenhuma das idades, então o algoritmo reconhece como a idade da vítima a primeira idade encontrada no texto. Isto ocorre por ter-se identificado que, geralmente, os primeiros dados pessoais presentes em uma notícia, como a idade, referem-se a vítima.

Após a execução dos dois passos descritos, o algoritmo extrai do texto a idade da vítima, caso esta tenha sido citada no texto da notícia.

3.1.2.4 Extração da identidade de gênero e orientação sexual da vítima

Esta subseção expõe, em detalhes, todos os passos utilizados para obter a identidade de gênero e orientação sexual das vítimas de violência, utilizando as informações encontradas em

notícias. São dados necessários para compreender, por exemplo, quais grupos da comunidade LGBTQIA+ estão mais suscetíveis a violência. Dada a pluralidade da comunidade LGBTQIA+, este trabalho focou em encontrar as identidades de gênero e orientações sexuais mais comuns em textos notícias. É possível incluir no algoritmo a busca por outras palavras que representem indivíduos LGBTQIA+.

A partir do texto extraído da notícia, o algoritmo realiza os seguintes passos para a extração da identidade de gênero e orientação sexual da vítima:

- Passo 01: encontrar no texto as palavras chaves que indiquem a identidade de gênero e orientação sexual da vítima. A partir de um conjunto pré-definido de palavras, e da tokenização [5] do texto, o algoritmo realiza uma busca no texto tentando encontrar neste toda palavra pertencente ao conjunto, salva as palavras que encontra e conta sua ocorrência cada vez que é encontrada. O conjunto de palavras a serem buscadas é constituído por dois subconjuntos, um referente a identidades de gênero e outro referente a orientações sexuais. Estes dois conceitos citados são distintos, mas isolados ou em conjunto podem definir uma pessoa como LGBTQIA+ já que um mesmo indivíduo pode possuir identidade e orientação divergentes das normativas. O conjunto de identidades de gênero buscados no texto é formado por: 'travesti', 'transgênero', 'trans', 'transexual', 'não-binário'. Caso as palavras 'transgênero', 'trans', 'transexual' sejam encontradas, verifica-se se a palavra anterior a elas são 'homem' ou 'mulher', para poder armazenar a informação completa sobre a identidade de gênero. O conjunto de orientações sexuais buscadas no texto é composto pelas palavras: 'lésbica', 'gay', 'bissexual', 'homossexual', 'pansexual', 'assexual'. Após percorrer o texto, têm-se uma tupla[6] de vetores, constituída por dois conjuntos distintos de palavras, representando as identidades de gênero e orientações sexuais encontradas e suas respectivas frequências; cada vetor pode conter mais de uma palavra ou mesmo estar vazio;

- Passo 02: obter a identidade de gênero e orientação sexual da vítima, caso tenham sido encontradas no texto. Para cada vetor da tupla retornada pelo Passo 01, ordena-se suas palavras encontradas pela ocorrência da mais frequente para a menos frequente. Desta forma, tanto no vetor de identidades de gênero encontradas, quanto no vetor de orientações sexuais encontradas, a primeira palavra do vetor, após o vetor ordenado, será a mais frequente no texto. Como se observou que a identidade e orientação da vítima geralmente são as mais citadas em textos de notícias, assumiu-se que a identidade e orientação mais frequentes referem-se como as da vítima e são estas as extraídas. Assim, após a ordenação dos vetores, o primeiro elemento do vetor de identidades de gênero é extraído como a

[5] Segundo Torres, Zaina e Almeida (2012) "o processo de tokenização pode ser definido como um processo de análise léxica, que analisa uma entrada de linhas de caracteres e gera uma sequência de símbolos. Durante a tokenização é necessário remover alguns caracteres indesejados, como sinais de pontuação, separação silábica, marcações especiais e números, os quais, isoladamente fornecem pouca informação".

[6] Uma tupla é uma estrutura de dados em programação que permite armazenar um conjunto ordenado de elementos heterogêneos.

identidade da vítima e, da mesma forma, o primeiro elemento do vetor de orientações sexuais é extraído como sendo a orientação da vítima.

Ao final da execução dos dois passos descritos, tem-se extraído a identidade de gênero e orientação sexual da vítima, caso estejam presentes no texto.

3.1.2.5 Extração da localidade em que a vítima sofreu violência

Nesta subseção, são abordadas, detalhadamente, as extrações dos dados, a partir de notícias, que compõem a localidade em que vítimas LGBTQIA+ sofreram violência. Essa extração obtém os dados do bairro, cidade, estado e região do Brasil, quando possível.

Para extração dos dados de localidade, há um pré-processamento, utilizando mineração de texto, para identificar palavras referentes a localidades no texto:

- Primeiramente, assim como no processo de extração do nome da vítima, é carregado o módulo *pt_core_news_lg* da biblioteca *SpaCy*. O que difere neste momento é que o rótulo "LOC"[7] é utilizado para extrair identidades do texto identificadas como localidade. Por meio deste processo, busca-se identificar no texto todas as palavas que são referentes a localidades;
- Após obter todas as palavras que se referem a lugares, o conjunto de dados é filtrado para não conter nenhuma palavra repetida.

Depois de extrair palavras referentes a localidade no texto, o primeiro dado a ser extraído é o bairro. Este dado é importante para se identificar regiões violentas para LGBTQIA+ dentro de um município. A extração do bairro de um texto de uma notícia dá-se pelos seguintes passos:

- Passo 01: retornar, a partir do conjunto de palavras que representam localidades, somente aquelas que são mencionadas no texto após a palavra bairro. O algoritmo procura pela palavra 'bairro' no texto e verifica se, 30 caracteres após a ocorrência desta palavra, quais das palavras que representam localidades estão contidas neste intervalo;
- Passo 02: contar a quantidade de ocorrência das palavras encontradas logo após a ocorrência da palavra 'bairro' retornadas pelo Passo 01. Para cada palavra retornada pelo passo anterior, percorre-se o texto e contabiliza-se quantas vezes a mesma foi citada;
- Passo 03: ordenar as palavras retornadas pelo Passo 01 das mais frequentes para as menos frequentes, a partir das frequências retornadas pelo Passo 02;

[7] LOC é o rótulo atribuído por modelos de processamento de linguagem treinados para identificar entidades nomeadas em um texto e indica a presença de nomes de lugares ou locais geográficos. Por meio do aprendizado de máquina, esses modelos conseguem reconhecer automaticamente referências a cidades, países, estados, endereços e outros elementos que descrevem localizações no texto, auxiliando na extração de informações geográficas.

- Passo 04: retornar a palavra mais frequente obtida pelo Passo 03. O algoritmo assume que a palavra mais frequente no texto, ocorrida após a palavra 'bairro', representa o bairro em que a violência ocorreu. Isto foi feito após se observar que, geralmente, quando o bairro é citado, costuma ser citado mais vezes ao longo do texto. E, caso existam citações de outros bairros, geralmente o mais citado é o em que o ato violento ocorreu.

Tendo também como entrada as palavras referentes a localidade encontradas no texto da notícia, o segundo dado a ser extraído trata-se da cidade em que vítima sofreu violência. Este dado tem como objetivo auxiliar a compreender quais cidades possuem altos índices de violência contra LGBTQIA+ no Brasil. O algoritmo de extração de cidades constitui-se nos seguintes passos:

- Passo 01: encontrar, a partir do conjunto de palavras referentes a localidade, as palavras que realmente são nomes de cidades brasileiras. Para cada palavra do conjunto de entrada, realiza-se uma busca binária[8] em um arquivo que contém o nome de todas cidades, junto com os seus respectivos códigos de UF, e armazena-se o nome da cidade e o seu código quando encontrada;
- Passo 02: contar a ocorrência dos nomes de cidades encontrados no Passo 01 em todo o texto. A partir dos nomes de cidades encontrados no Passo 01, para cada um deles, realiza-se uma busca no texto contabilizando a quantidade de ocorrência dos mesmos;
- Passo 03: ordenar os nomes das cidades encontrados pelo Passo 01, por meio das ocorrências contabilizadas pelo Passo 02, dos mais frequentes para os menos frequentes;
- Passo 04: retornar o nome da cidade mais citado no texto junto com o código UF encontrado no Passo 01. A partir do Passo 03, identifica-se a cidade mais citada no texto e a retorna em conjunto com seu respectivo código UF. Observou-se que, geralmente, a cidade mais citada no texto trata-se da cidade em que o ato violento ocorreu.

A partir dos passos anteriores, busca-se extrair a cidade em que ocorreu uma violência contra um LGBTQIA+ brasileiro em uma notícia. Quando a cidade é extraída e em conjunto com seu código da UF, pode-se então obter o estado e a região do Brasil em que esta se encontra. Para obter essas informações, tem-se um arquivo contendo o nome de todos os estados brasileiros, seus respectivos códigos UF e suas respectivas regiões. A partir do código UF retornado em conjunto com o nome da cidade, realiza-se uma busca neste arquivo e obtém-se as informações sobre qual estado e qual a região do Brasil em que o ato presente na notícia ocorreu.

[8] Segundo Santamarina (2019), busca binária "é um algoritmo de busca para encontrar um determinado valor dentro de um conjunto de dados ordenados. O algoritmo sucessivamente diminui o espaço de busca pela metade, até que o elemento seja encontrado ou então até que se constate que ele não existe nos dados".

Os processos descritos anteriormente, ao serem combinados, permitem a extração das informações sobre a localização em que uma vítima LGBTQIA+ sofreu violência, quando estas estão presentes no texto.

3.1.2.6 Extração dos tipos de violência sofridos pela vítima

Esta subseção aborda, em detalhes, como é realizado o processo de extração dos tipos de violência sofridos pela vítima. Como os tipos de violência são muitos, primeiramente analisou-se notícias de violência contra pessoas LGBTQIA+ e identificou-se os tipos mais presentes nestas notícias. Definiu-se e conceituou-se neste trabalho, como os quatro principais tipos de violência, os tipos: assassinato, agressão física, agressão verbal e privação de direitos. Em uma definição mais objetiva, estes conceitos de tipo de violência definem-se como:

- **assassinato**: consiste em identificar os casos em que a vítima foi assassinada. Para isto, busca identificar-se no texto as palavras: 'mort' (radical para encontrar palavras como morta e morto), 'assassinad' (radical para encontrar palavras como assassinado e assassinada) e 'corpo encontrado';
- **agressão física**: consiste em identificar os casos em que a vítima sofreu agressão física. Isto é feito por meio da busca dos seguintes termos no texto: 'soco', 'facada', 'chute', 'agredid'(radical para encontrar palavras como agredido e agredida) , 'ferimento', 'corte', 'agressão', 'lesão' e 'empurrad' (radical para encontrar palavras como empurrado e empurrada);
- **agressão verbal**: consiste em identificar os casos em que a vítima sofreu agressão verbal. Por meio da busca de palavras específicas no texto, consegue-se identificar a ocorrência de agressão verbal contra a vítima. As palavras buscadas são: 'humilhad' (radical para encontrar palavras como humilhado e humilhada), 'insultad' (radical para encontrar palavras como insultado e insultada), 'xingad' (radical para encontrar palavras como xingado e xingada) e 'ridicularizad' (radical para encontrar palavras como ridicularizado e ridicularizada);
- **privação de direitos**: consiste em identificar os casos em que a vítima foi privada de seus direitos. Para esta identificação, buscou-se no texto as palavras: 'privad' (radical para encontrar palavras como privado e privada), 'expuls' (radical para encontrar palavras como expulso e expulsa) e 'proibid' (radical para encontrar palavras como proibido e proibida).

A partir dos tipos de violência definidos e das palavras utilizadas para o reconhecimento de cada um deles, o algoritmo realiza os seguintes passos para extrair os tipos de violência:

- Passo 01: percorrer o texto busca cada palavra que define cada tipo de violência no texto. A partir de um conjunto de palavras que representam violências, contendo todas as palavras

que definem todos os tipos de violência, o algoritmo percorre o texto buscando encontrar as palavras presentes no conjunto;

- Passo 02: armazenar as violências identificadas no texto e os tipos de violência encontradas. Quando o algoritmo encontra alguma palavra do conjunto de palavras que representa violências, ele armazena a palavra encontrada e também armazena a qual tipo de violência esta pertence. Por exemplo, quando a palavra 'soco' é encontrada, armazena-se esta palavra e o seu tipo de violência que consiste em 'agressão física'.

A partir destes passos informados, os tipos de violência são identificados no texto e consegue-se levantar informações estatísticas sobre os tipos de violência, dentre os citados, mais acometidos contra a comunidade LGTQIA+ no aspecto geral e, também, quais os tipos de violência mais frequentes contra cada grupo constituinte da comunidade LGBTQIA+. Além dos tipos de violência definidos neste trabalho, consegue-se obter também informações estatísticas sobre as violências em si sofridas pela vítimas, por meio das palavras armazenadas que representam estas violências.

3.1.3 Integração de dados

Esta Subseção detalha o Passo 3 da arquitetura de funcionamento (vide Figura 3.1), que consiste no processo de integração dos dados retornados pelo Passo 2 .

Um conjunto de arquivos XML, onde cada arquivo contém todas as informações extraídas de uma notícia, como apresentado na Figura 3.3, é recebido como entrada. Para cada arquivo recebido, são percorridas as informações presentes no arquivo e elas são comparadas com as existentes em um arquivo XML único que serve como um banco de dados semi-estruturado. É importante ressaltar que este XML consiste no arquivo XML final gerado pelo processo de integração. Quando o processo de integração inicia, verifica-se se este arquivo já existe e, caso ele não exista, ele é então criado e os dados da primeira notícia a ser integrada são salvos; a partir destes dados salvos, os dados das novas notícias a serem armazenadas são então comparados. As informações comparadas com o arquivo que consiste no banco de dados semi-estruturado são, especificamente: mês e ano da notícia, nome da vítima, idade da vítima, orientação sexual da vítima, identidade de gênero da vítima, estado em que a vítima sofreu violência e região do Brasil em que a vítima sofreu violência. A biblioteca *xml.etree.ElementTree* da linguagem Python é responsável por recuperar os dados presentes no banco de dados e do arquivo XML com as informações de uma nova notícia. Para cada notícia a ser armazenada, a mesma é comparada com todas as notícias presentes no banco de dados. Para cada notícia do banco de dados, verifica-se se todos os dados da notícia já armazenada estão presentes na notícia a ser armazenada ou se todos os dados da notícia a ser armazenada estão presentes nos dados da notícia já armazenada. Se uma destas afirmações for verdadeira, isto indica que os dados da notícia a serem armazenados são dados de um caso de violência já salvos. Caso a notícia contenha algum dado a mais dos

que os já armazenados, este dado então é inserido no banco de dados no mesmo conjunto de dados que representa o caso de violência armazenado. Caso não haja similaridade com os dados já armazenados oriundos de alguma página, um identificador único é criado para identificar as informações daquela notícia e o objeto XML contido no arquivo é armazenado no banco. Este processo ocorre para que não haja duplicidade entre os dados armazenados; desta forma, cada objeto XML armazenado, contendo seu respectivo identificador, representa um único caso de violência específico. O banco de dados contendo todos os dados extraídos e integrados das notícias coletadas no Passo 1 serve então como entrada do Passo 4 (vide Figura 3.1), como já mencionado.

3.2 Interfaces da ferramenta resultante da estratégia proposta

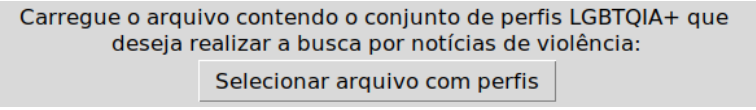
Nesta seção, são apresentadas as interfaces da ferramenta resultante da estratégia proposta neste trabalho. A Subseção 3.2.1 aborda as interfaces referentes à coleta de páginas de notícias sobre violência contra LGBTQIA+ brasileiros, já que corresponde ao início da estratégia proposta e envolve interação com o usuário. E a Subseção 3.2.2 aborda as interfaces da plataforma de visualização de dados *Qlik Cloud*, já que representa o final da estratégia proposta e também envolve interação com o usuário.

3.2.1 Interface da coleta de páginas de notícias sobre violência contra LGBTQIA+ brasileiros

Esta subseção apresenta as telas de interação com o usuário para realização da coleta de páginas de notícias contendo casos de violência contra a comunidade LGBTQIA+ brasileira.

O passo inicial do algoritmo de coleta de páginas é apresentado na Figura 3.4. Nesta tela, o usuário precisa escolher o arquivo contendo os perfis LGBTQIA+ a serem encontrados em notícias de casos de violência. Este arquivo é criado pelo usuário, conforme explicado na Subseção 3.1.1.

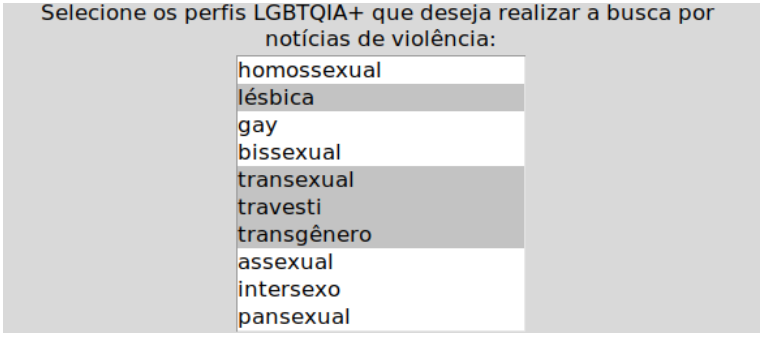
Figura 3.4 – Escolha do arquivo contendo perfis LGBTQIA+



Fonte: elaborado pela autora

Uma vez escolhido o arquivo contendo os perfis LGBTQIA+ a serem encontrados em notícias de casos de violência, o usuário precisa então selecionar os perfis LGBTQIA+ a serem buscados em notícias de casos de violência, como pode ser visto na tela da Figura 3.5. Os perfis exibidos ao usuário para seleção são os mesmos presentes no arquivo que ele selecionou.

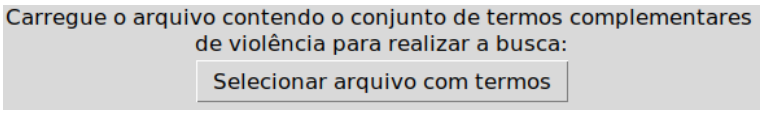
Figura 3.5 – Seleção dos perfis LGBTQIA+



Fonte: elaborado pela autora

Logo após, o usuário escolhe o arquivo contendo os termos de violência que deseja associar a cada perfil LGBTQIA+ selecionado para realizar a busca por notícias de violência, conforme pode ser visualizado na Figura 3.6. Este arquivo também é criado pelo usuário, conforme explicado na Subseção 3.1.1.

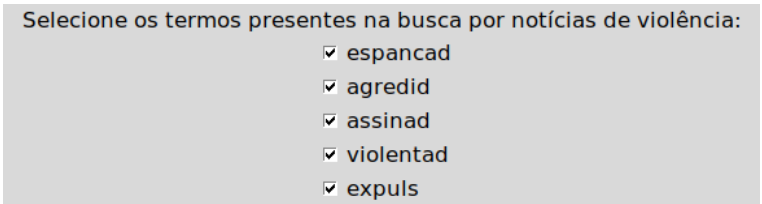
Figura 3.6 – Seleção do arquivo contendo termos de violência



Fonte: elaborado pela autora

Após escolher o arquivo contendo os termos de violência que deseja associar a cada perfil LGBTQIA+ selecionado, o usuário precisa então selecionar os termos para realizar a busca por notícias, como pode ser observado na Figura 3.7. Os termos de violência exibidos ao usuário para seleção são os mesmos presentes no arquivo que ele selecionou.

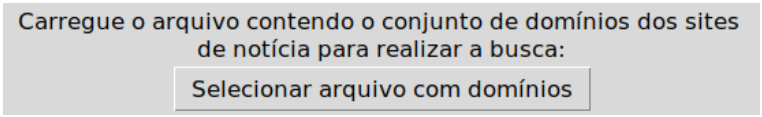
Figura 3.7 – Seleção dos termos de violência



Fonte: elaborado pela autora

A próxima ação do usuário é escolher o arquivo contendo os domínios dos sites em que se deseja encontrar as notícias de violência, como pode ser visto na Figura 3.8. Isto é necessário para que a coleta de páginas seja realizada nestes domínios especificados pelo usuário, conforme explicado na Subseção 3.1.1.

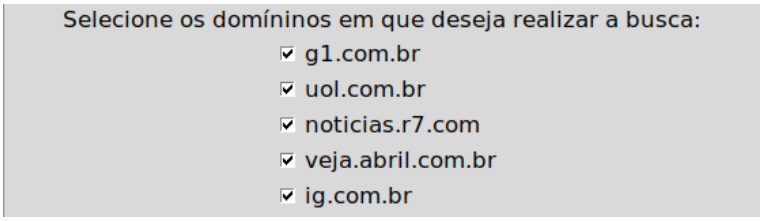
Figura 3.8 – Seleção do arquivo contendo domínios de sites de notícias



Fonte: elaborado pela autora

Depois de escolher o arquivo contendo os domínios dos sites, o usuário deve então selecionar os domínios dos sites em que se deseja encontrar as notícias de violência, como pode ser visualizado na Figura 3.9. Os domínios dos sites exibidos ao usuário para seleção são os mesmos presentes no arquivo que ele selecionou.

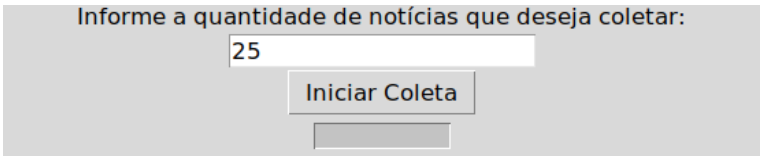
Figura 3.9 – Seleção dos domínios de sites de notícias



Fonte: elaborado pela autora

Em seguida, o usuário deve especificar a quantidade de notícias que devem ser retornadas pelo algoritmo coletor, como pode ser observado na Figura 3.10. Após o usuário informar a quantidade, tem-se então todas as informações necessárias para a realização da coleta.

Figura 3.10 – Especificação da quantidade de notícias a serem retornadas



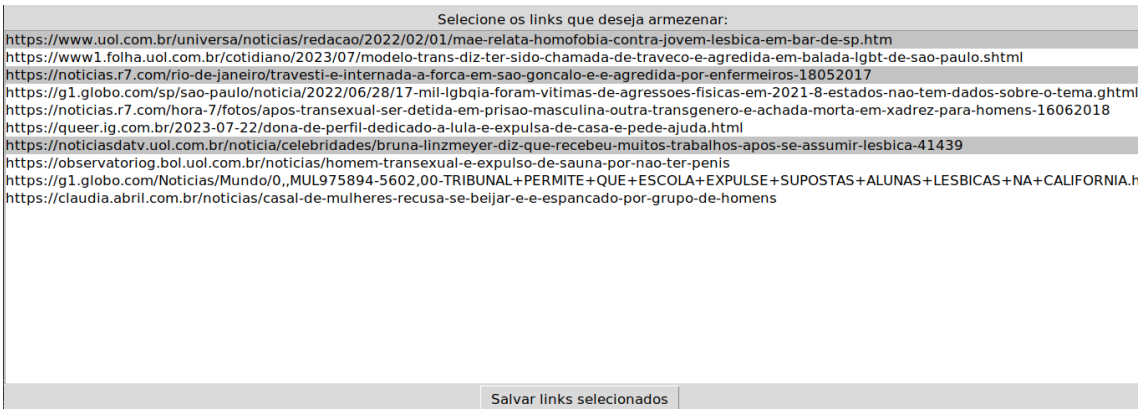
Fonte: elaborado pela autora

O usuário deve então clicar no botão "Iniciar coleta", que pode ser visualizado na 3.10. Após esta ação, a coleta de notícias de violência é devidamente iniciada. Enquanto a coleta acontece, a barra de *loading* abaixo do botão indica ao usuário que a coleta ainda está ocorrendo.

Após a coleta das páginas pelo algoritmo, os *links* coletados são, então, apresentados ao usuário e este deve então analisar e selecionar os *links* que acredita serem referentes a casos violência contra LGBTQIA+ brasileiros, conforme pode ser visto na tela da Figura 3.11.

Depois de selecionar os *links*, o usuário deve então clicar no botão 'Salvar *links* selecionados', como pode ser também observado na Figura 3.11. Por fim, após a seleção dos *links*

Figura 3.11 – Links de páginas de notícias encontradas pelo coletor



Fonte: elaborado pela autora

pelo usuário, os mesmos são armazenados em um arquivo que consiste no conjunto de *links* fornecidos como entrada para o Passo 2 da arquitetura de funcionamento (vide Figura 3.1).

3.2.2 Interfaces da plataforma de visualização de dados *Qlik Cloud*

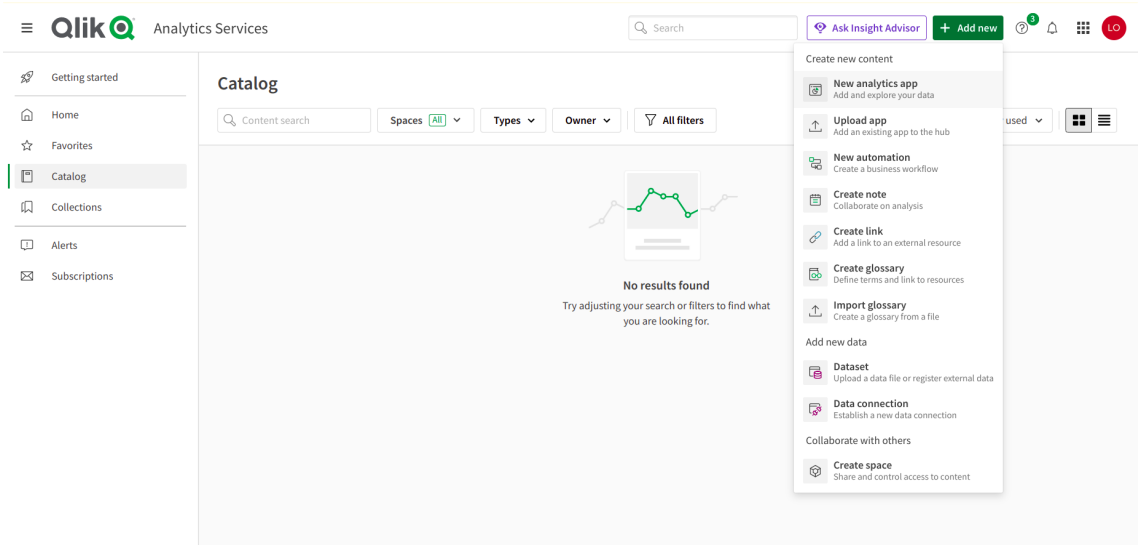
Esta subseção aborda as interfaces da plataforma online de visualização de dados *Qlik Cloud*. Apesar de ser uma plataforma em que se necessita pagar por sua utilização, este trabalho optou por utilizar seus trinta dias de teste grátis para demonstrar a visualização dos dados obtidos pelo Passo 3 da arquitetura de funcionamento (vide Subseção 3.1). Como a *Qlik Cloud* é uma plataforma que fornece muitos recursos, optou-se por exibir apenas as interfaces dos recursos necessários para a visualização dos dados de violência contra LGBTQIA+ brasileiros.

Para a utilização da plataforma *Qlik Cloud* deve-se, primeiramente, criar uma conta na plataforma. Na tela inicial, a seção *Catalog* do menu principal deve ser acessada. No canto superior direito, o botão *Add new* deve ser clicado e, dentro das opções disponíveis, a opção *New analytics app* deve ser escolhida. A tela inicial pode ser observada na Figura 3.12.

Após a escolha da opção *New analytics app*, uma tela é aberta para a criação de uma nova aplicação dentro da plataforma. Neste momento, deve-se definir um nome para a aplicação que conterá todos os gráficos referentes aos dados de violência contra LGBTQIA+ brasileiros. Após informar o nome, mantenha a opção *Open app* selecionada e basta clicar em *Create* para criar a aplicação. A tela de criação de uma nova aplicação pode ser vista na Figura 3.13.

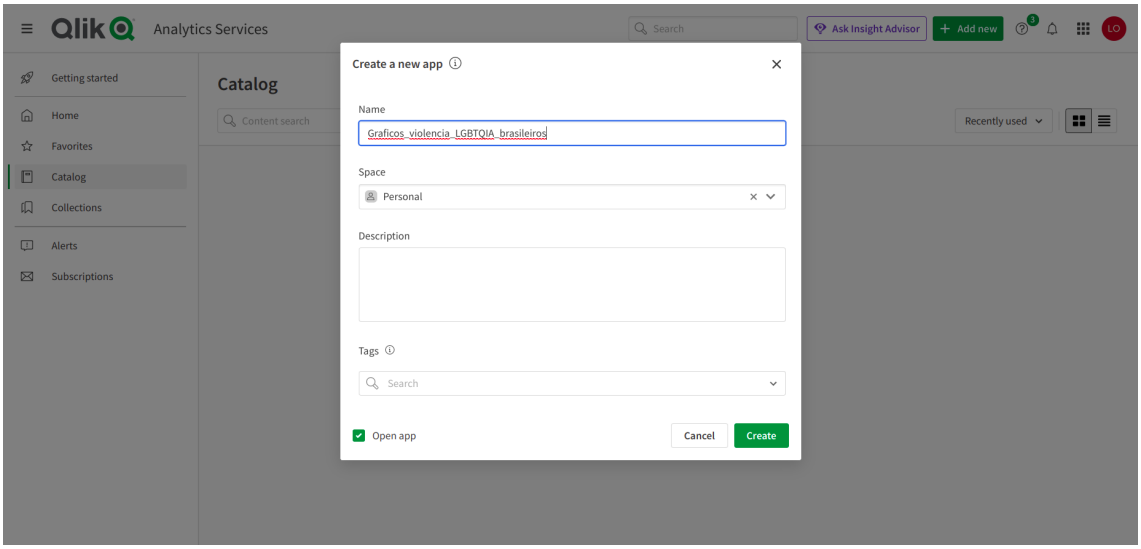
Depois de criar aplicação, é necessário carregar os dados de entrada da aplicação. Neste trabalho, os dados de entrada da aplicação consistem no arquivo XML, que representa o banco de dados semi-estruturado, com os dados extraídos de todas as notícias. Após criar no botão *Create*, conforme visto na Figura 3.13, a aplicação estará aberta. Neste momento, é preciso clicar em *Files and other sources* para carregar os dados da aplicação, que pode ser observado na Figura 3.14.

Figura 3.12 – Tela inicial da plataforma *Qlik Cloud*



Fonte: elaborado pela autora

Figura 3.13 – Tela de criação de uma nova aplicação

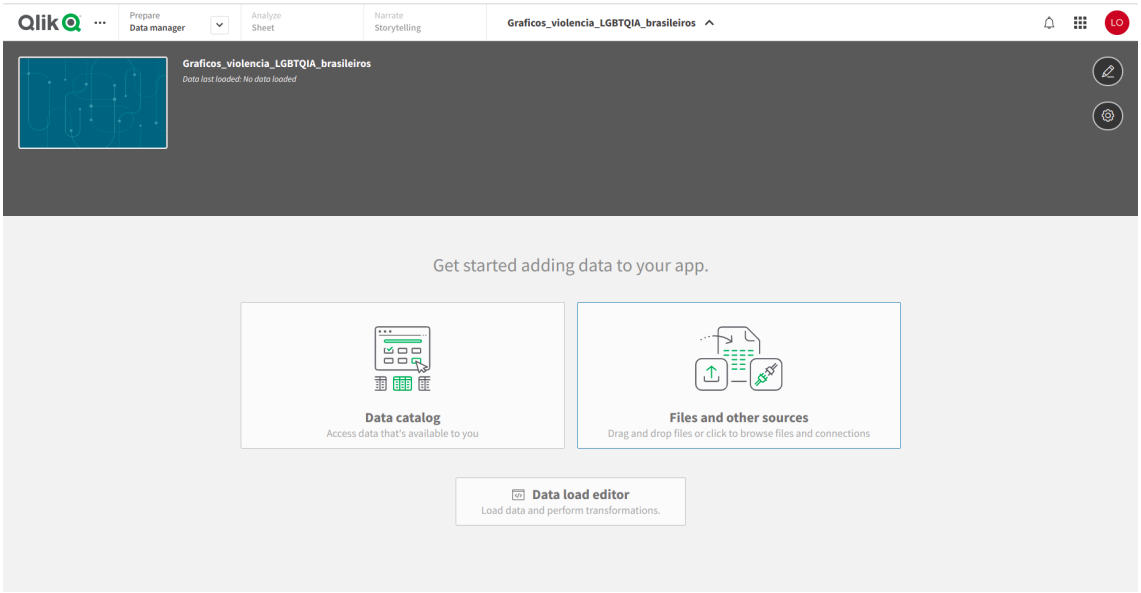


Fonte: elaborado pela autora

Logo após clicar em *Files and other sources*, uma tela para adicionar um arquivo ou base de dados como dado de entrada será exibida. Basta clicar no botão *Drop a file here or click to select a file* para escolher o arquivo XML que consiste no banco de dados semi-estruturado gerado pela estratégia descrita neste trabalho. A tela e o botão podem ser vistos na Figura 3.15.

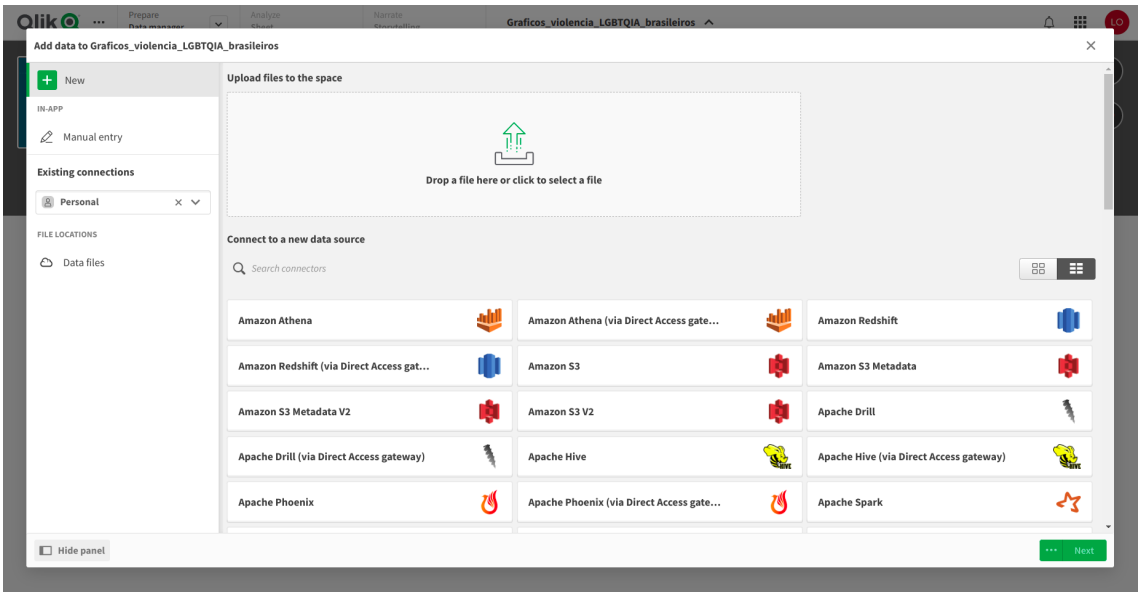
Após carregar o arquivo XML contendo os dados extraídos das notícias, a aplicação realiza então algumas associações de dados e cria algumas tabelas. Neste momento, basta clicar no botão *Next* presente no canto inferior direito. A tela contendo as tabelas criadas e botão podem ser vistos na Figura 3.16.

Figura 3.14 – Tela *upload* de dados para a aplicação



Fonte: elaborado pela autora

Figura 3.15 – Tela *upload* de arquivo para a aplicação

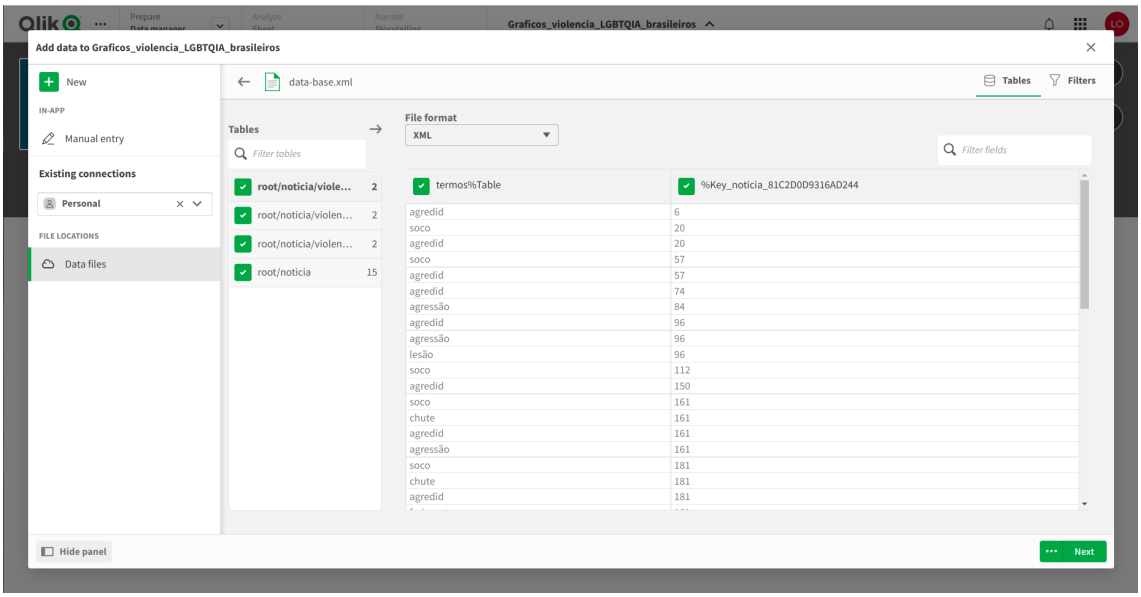


Fonte: elaborado pela autora

Depois de clicar em *Next*, a aplicação exibe as tabelas criadas como conjuntos representados por círculos na tela. É necessário conectar esses conjuntos para a devida associação dos dados entre as tabelas. Os conjuntos criados a partir do XML de entrada podem ser observados na Figura 3.17.

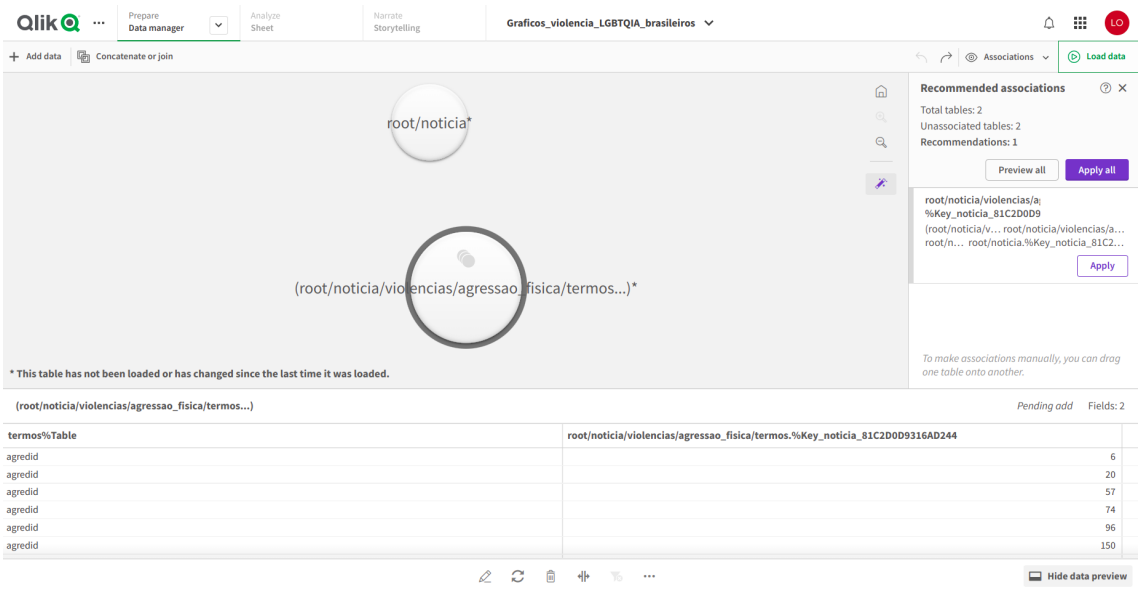
É preciso conectar os conjuntos presentes na Figura 3.17; para isto, basta arrastar um em direção ao outro, seus dados são associados e a visualização destes conjuntos conectados é

Figura 3.16 – Tela de dados obtidos a partir de um arquivo de entrada



Fonte: elaborado pela autora

Figura 3.17 – Conjuntos criados pela aplicação a partir do XML de entrada



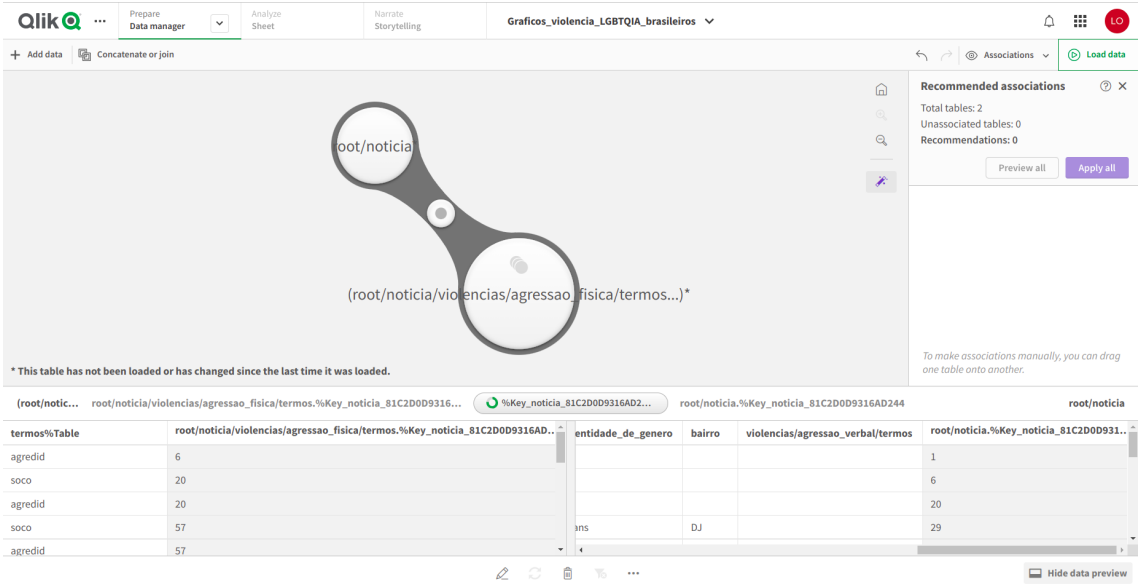
Fonte: elaborado pela autora

exibida na Figura 3.18.

Em sequência a associação de dados, é preciso que a aplicação carregue todas essas associações para geração dos gráficos. Para isto, basta clicar em *Load data* no canto superior da tela, conforme apresentado na Figura 3.18.

A próxima ação a ser tomada depois de carregar os dados é ir para o painel que permite a criação dos gráficos para a visualização dos dados. Após finalizar o carregamento dos dados,

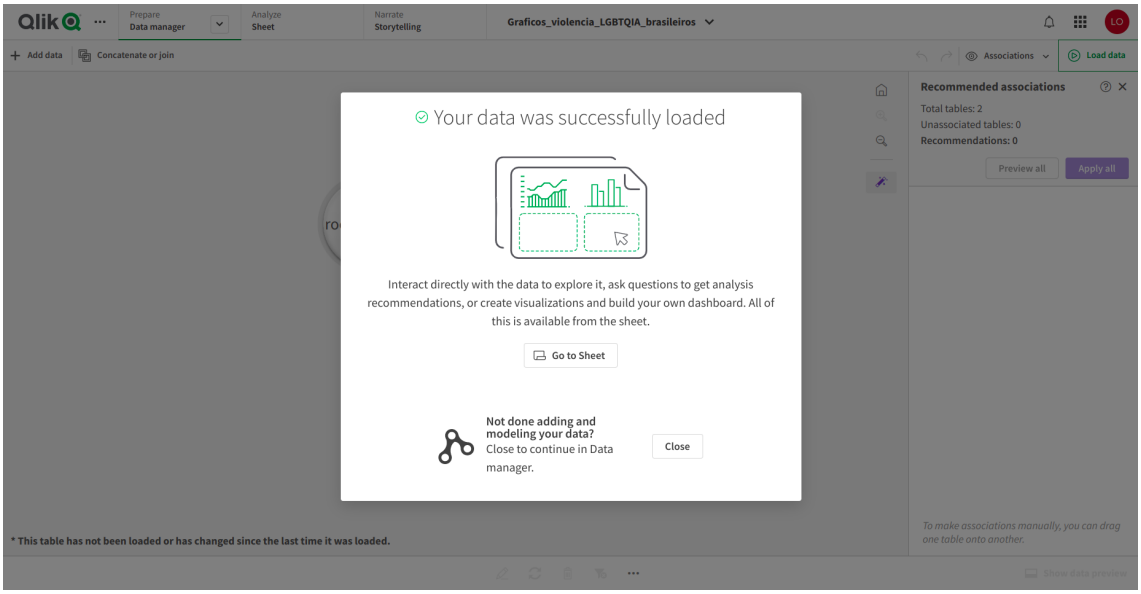
Figura 3.18 – Conexão entre os conjuntos de dados criados



Fonte: elaborado pela autora

a aplicação exibe uma tela de sucesso, conforme pode ser visto na Figura 3.19. Basta clicar no botão *Go to Sheet* presente nesta tela para abrir o painel.

Figura 3.19 – Tela de sucesso do carregamento dos dados

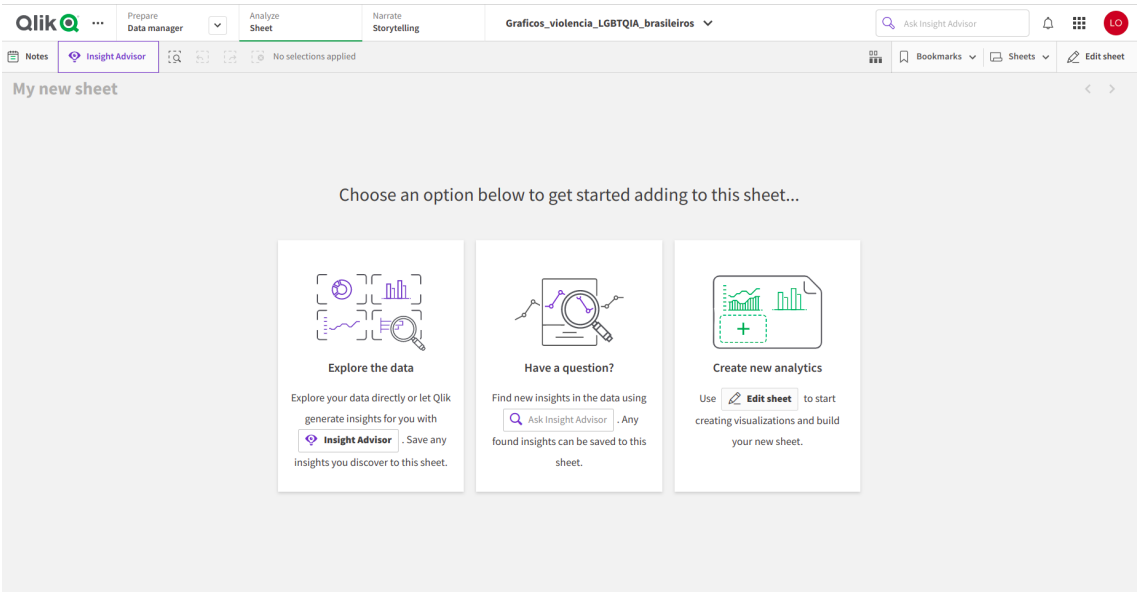


Fonte: elaborado pela autora

A tela inicial do painel apresenta algumas opções para se trabalhar com os dados carregados e associados dentro da aplicação *Qlik Cloud*, conforme pode ser visto na Figura 3.20. Este trabalho tem como objetivo final a visualização, por meio de gráficos, de dados estatísticos sobre

violências contra LGBTQIA+ brasileiros, obtidos a partir de sites de notícias. Dado este foco, a opção *Create new analytics* deve ser escolhida na tela inicial.

Figura 3.20 – Tela de seleção da opção para criação de visualizações de dados



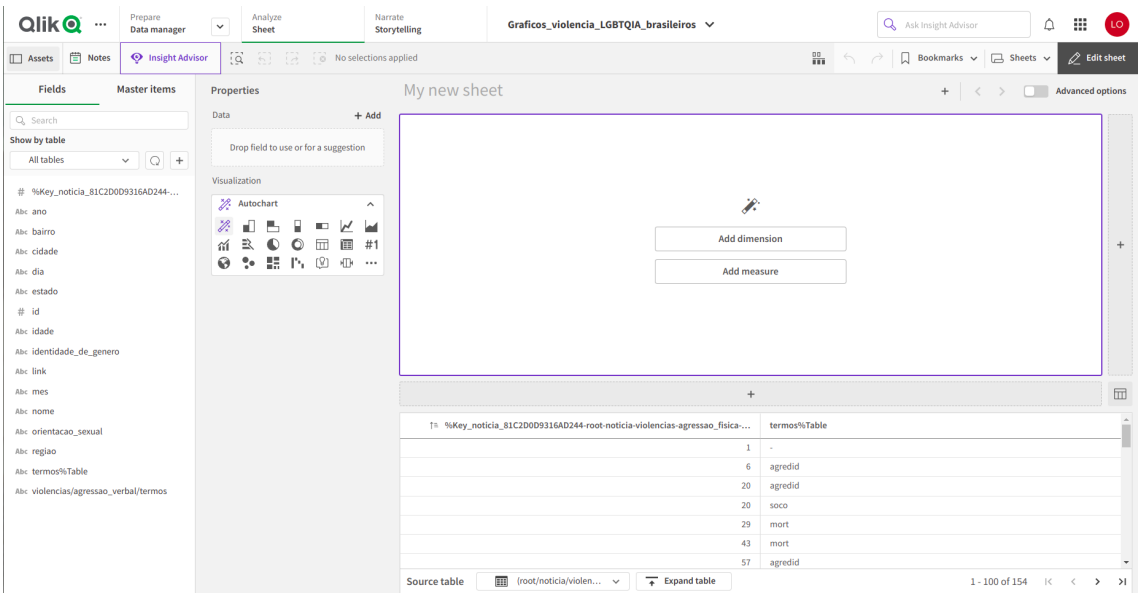
Fonte: elaborado pela autora

Após clicar na opção *Create new analytics*, o painel principal para criação de visualizações de dados é aberto e este painel é chamado de *sheet*. Uma mesma aplicação pode possuir mais de um painel, e cada painel pode possuir um ou vários gráficos de visualização de dados; isto permite uma melhor organização das visualizações de dados criados na plataforma. Um exemplo de *sheet* pode ser visualizado na Figura 3.21.

A Figura 3.21 possui a interface principal para a geração dos gráficos de informações estatísticas sobre violência contra LGBTQIA+ brasileiros. A coluna principal à esquerda representa os dados que foram extraídos e associados do arquivo XML de entrada. Para cada *tag* do arquivo de entrada, a aplicação *Qlik Cloud* interpreta como um dado distinto que pode gerar associações e gráficos distintos. Logo ao lado direito da coluna, na seção *Visualization*, é possível escolher os tipos de gráficos a serem gerados e outros elementos de interação com os gráficos, como filtros, por exemplo.

Como exemplo de visualização dos dados, a Figura 3.22 apresenta um gráfico gerado a partir de dados de dados de violência contra a população LGBTQIA+ brasileira, extraídos de notícias utilizando a estratégia descrita neste trabalho (vide Subseção 3.1). Para a criação deste gráfico, escolheu-se a opção região como o dado a ser apresentado em um gráfico. A *tag* região do arquivo XML tem como valor 5 possíveis regiões: Norte, Nordeste, Centro-Oeste, Sudeste, Sul. Para as notícias que não possuíam a *tag* região no banco de dados semi-estruturado, a associação de dados deixa este campo em branco para estas notícias dentro da aplicação *Qlik Cloud*. Adicionou-se, também, uma visualização de filtro para filtrar entre as regiões presentes

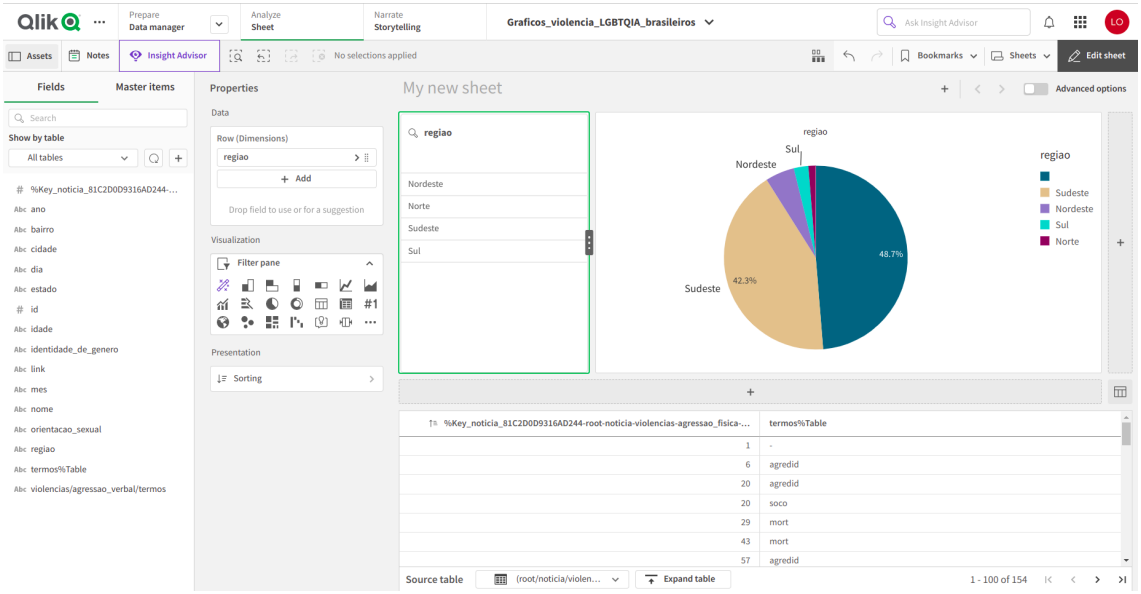
Figura 3.21 – Painel principal para criação de visualizações de dados



Fonte: elaborado pela autora

no gráfico a ser exibido. A tela com o gráfico, que contabiliza a quantidade de casos de violência por região do Brasil, e o filtro por região podem ser vistos na Figura 3.22.

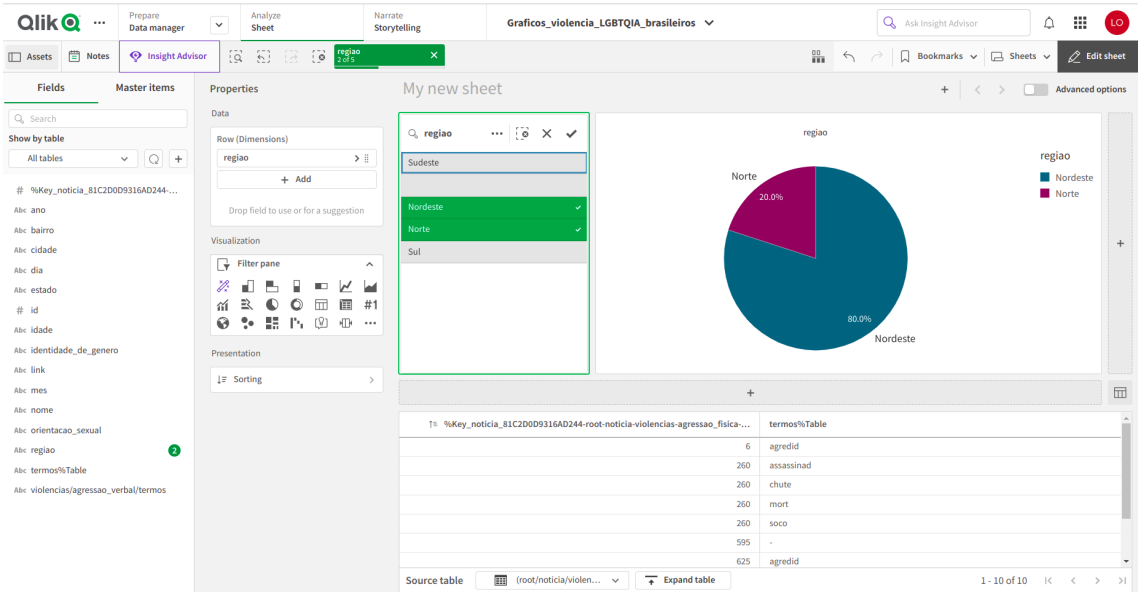
Figura 3.22 – Exemplo de gráfico criado a partir dos dados de violência por região do Brasil



Fonte: elaborado pela autora

O gráfico é totalmente interativo, permitindo filtrar ao clicar nas regiões presentes no filtro (elemento que pode ser observado logo a esquerda do gráfico na Figura 3.22) e ter informações mais detalhadas ao clicar nas distintas regiões presentes no gráfico. Um exemplo de visualização utilizando o filtro pode ser observada na Figura 3.23.

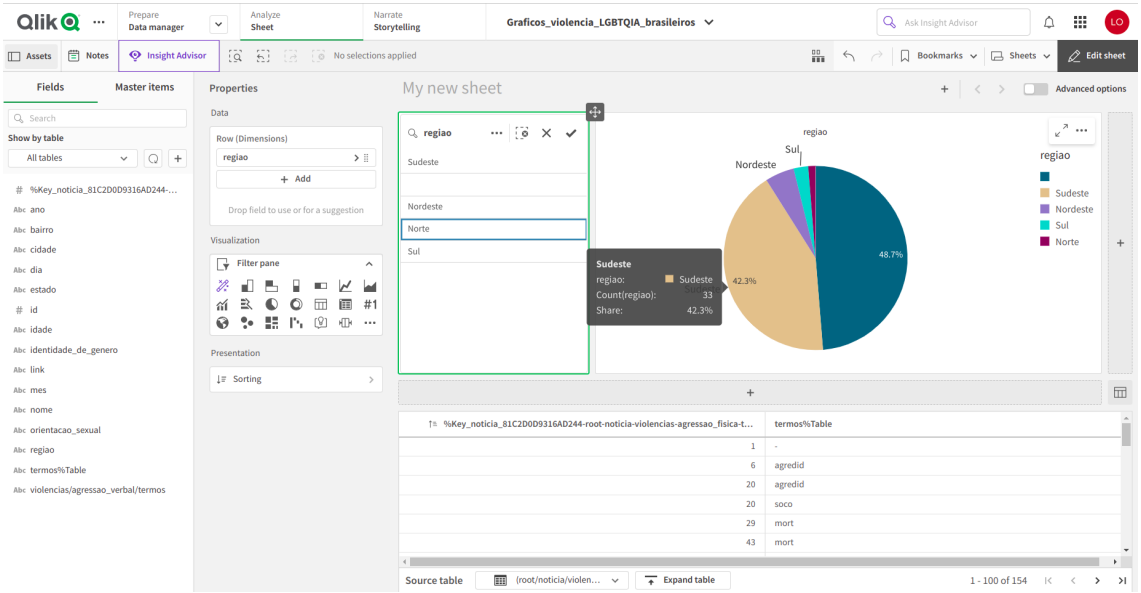
Figura 3.23 – Visualização de dados com filtro ativo



Fonte: elaborado pela autora

Outro exemplo de visualização de interação pode ser observado na Figura 3.24, onde é possível, ao deixar o cursor do *mouse* em cima de uma das regiões do mapa, obter mais informações sobre aquela região do gráfico, como: a quantidade de casos contabilizadas na região e o quanto esses casos representam em porcentagem em relação a todos os casos de todas as regiões.

Figura 3.24 – Informações presentes em partes dos gráficos gerados

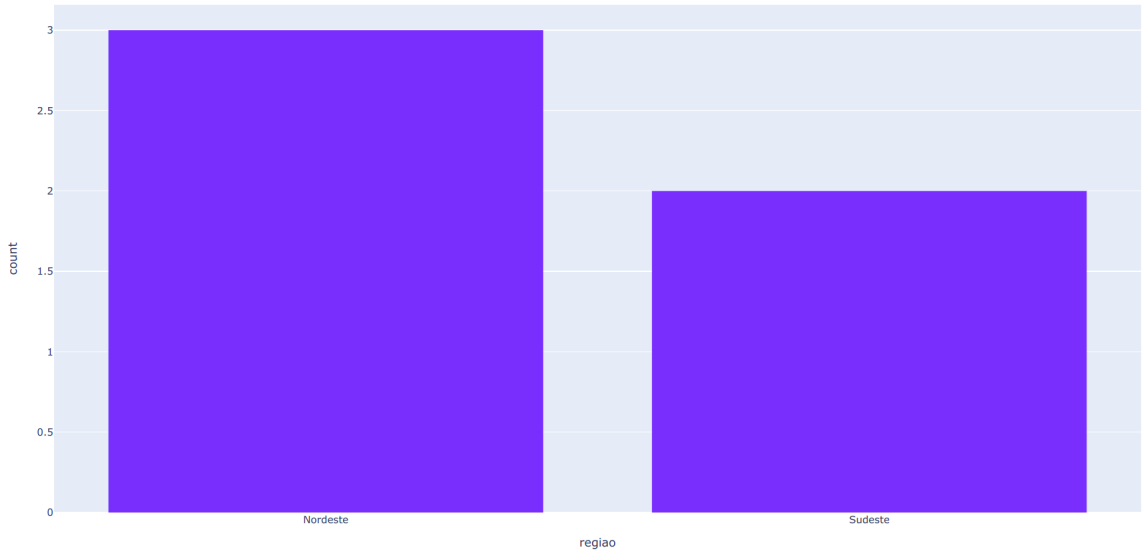


Fonte: elaborado pela autora

Com os exemplos apresentados da utilização e dos resultados obtidos com a plataforma

Qlik Cloud, é possível afirmar que se pode visualizar e inferir informações a partir dos dados estatísticos gerados a partir da estratégia elaborada neste trabalho, sendo que a visualização destes dados não fica restrita a utilização da plataforma *Qlik Cloud*; outras aplicações, que aceitem arquivos XML como entrada, podem também ser utilizadas. Outro meio é a possibilidade de desenvolver aplicações específicas para visualização dos dados gerados por este trabalho; um exemplo pode ser visto na Figura 3.25 onde, por meio de uma biblioteca *Python* de geração de gráficos *plot.express* somada a biblioteca de consultas XML *xml.etree.Element*tree, é possível também criar gráficos a partir do arquivo XML gerado pela estratégia. A Figura 3.25 apenas exemplifica um gráfico, populado com poucos dados, também exibindo casos de violência contabilizados em regiões do Brasil.

Figura 3.25 – Exemplo de gráfico gerado por um algoritmo próprio



Fonte: elaborado pela autora

4 Experimentação Prática

Neste capítulo, são apresentados e analisados os experimentos de validação da primeira versão funcional da estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira, seguindo a arquitetura proposta na Figura 3.1. A Seção 4.1 descreve a métrica utilizada para avaliar a eficácia da estratégia. A Seção 4.2 descreve todos os experimentos realizados. Por fim, a 4.3 apresenta a visualização dos resultados obtidos pela estratégia definida neste trabalho.

4.1 Métrica de Avaliação

Para avaliar todos os experimentos realizados com a estratégia definida neste trabalho, optou-se por utilizar a métrica de precisão. A precisão de um algoritmo pode ser definida como uma métrica que mede a taxa de acertos do algoritmo em relação ao número total de instâncias avaliadas. Ou seja:

$$\text{precisão} = \frac{\text{Número de acertos}}{\text{Número total de instâncias}}$$

4.2 Descrição dos Experimentos

A estratégia proposta e desenvolvida neste trabalho é composta de quatro módulos principais, como pode ser observado na arquitetura de funcionamento (vide Figura 3.1), sendo estes: a coleta semiautomática de páginas *Web* contendo notícias sobre violência contra a população LGBTQIA+ brasileira, extração de dados a partir destas páginas, integração destes dados extraídos e geração e apresentação dos dados. Tendo estes módulos como base, procurou-se experimentar e avaliar a precisão da coleta, extração e integração dos dados para avaliar a precisão de cada módulo. A apresentação dos dados constitui o objetivo final da presente estratégia e sua validação dá-se pela observação dos resultados obtidos.

Para que fosse possível experimentar a primeira versão funcional da estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira, primeiramente realizou-se a coleta semiautomática de páginas de notícias de violência contra LGBTQIA+ brasileiros. Os parâmetros necessários para a coleta foram:

- **Perfis LGBTQIA+**: definiu-se um arquivo texto contendo, em cada linha, as palavras presentes na Tabela 4.1 que representam os perfis LGBTQIA+ utilizados na busca por notícias de violência. Cada palavra da tabela contempla um perfil que compõe a comunidade LGBTQIA+.

Tabela 4.1 – Perfis LGBTQIA+ utilizados na busca por notícias de violência contra LGBTQIA+

Perfis LGBTQIA+
homossexual
lésbica
gay
bissexual
transexual
travesti
transgênero
assexual
intersexo
pansexual

Fonte: elaborado pela autora

- **Termos de violência**: definiu-se, em outro arquivo texto, as palavras presentes na Tabela 4.2 que são os termos de violência utilizados para a coleta de notícias. Cada termo representa o radical de uma palavra que caracteriza algum tipo de violência e foi armazenado em uma linha distinta do arquivo.

Tabela 4.2 – Termos de violência utilizados na busca por notícias de violência contra LGBTQIA+

Termos de violência
espancad
agredid
violentad
expuls
assassinad
proibid

Fonte: elaborado pela autora

- **Domínios de sites de notícias**: definiu-se, em outro arquivo texto distinto, as palavras presentes na Tabela 4.3 que constituem os domínios dos sites de notícias utilizados para a coleta de notícias. Foram escolhidos os principais sites de notícias do Brasil e cada domínio foi armazenado em uma linha distinta do arquivo.

Tabela 4.3 – Domínios dos sites de notícias utilizados na busca por notícias de violência contra LGBTQIA+

Domínios dos sites
g1.com.br
uol.com.br
noticias.r7.com
veja.abril.com.br
ig.com.br

Fonte: elaborado pela autora

- **Quantidade de notícias a serem buscadas**: definiu-se a quantidade de notícias a serem retornadas pelo coletor a cada utilização do coletor para a coleta de notícias de violência contra LGBTQIA+ brasileiros.

Os parâmetros citados anteriormente foram os utilizados para a realização da coleta páginas *Web* contendo notícias de violência contra LGBTQIA+ brasileiros. O buscador da *Google* apresentou algumas limitações em relação a consultas em um curto intervalo de tempo, muitas vezes suspendendo temporariamente o acesso ao seu recurso. Após perceber-se que até 100 páginas poderiam ser retornadas em um intervalo próximo de tempo, sem perder acesso ao recurso do buscador, foram realizados três processos de coleta, em intervalos de 30 minutos entre cada uma delas, buscando obter 100 páginas a cada coleta. Cada um dos três processos de coleta constituiu-se em executar o algoritmo quatro vezes, utilizando todos os parâmetros descritos anteriormente, onde todos os termos presentes em todos os arquivos foram inseridos na busca. Para cada processo de coleta forneceu-se, especificamente, o valor 25 a quantidade de notícias a serem retornadas por cada execução do algoritmo. Este valor foi estabelecido para não gerar uma sobrecarga de consultas em um curto espaço de tempo e ocasionar em algum bloqueio de consultas por parte do buscador utilizado. No total, sem levar em consideração os intervalos entre as execuções, foram realizadas 12 execuções do algoritmo de coleta com os parâmetros descritos, onde cada execução retornou 25 páginas de notícias. Foram obtidas 300 páginas de notícias no total em três processos de coletas distintos onde, cada processo, em intervalos de 30 minutos, retornou 100 páginas por meio da combinação de 4 execuções do algoritmo que retornaram 25 páginas cada uma.

A cada execução do algoritmo de coleta, para cada conjunto de 25 *links* de páginas de notícias retornadas, foram analisados todos os *links*, buscando identificar e selecionar aqueles que realmente correspondiam a páginas de notícias de violência contra LGBTQIA+ brasileiros, conforme descrito na Subseção 3.2.1. Este processo foi realizado para todas as 12 execuções do algoritmo e, das 300 páginas retornadas, 116 foram selecionadas como páginas válidas de violência contra LGBTQIA+ brasileiros. Das 116 páginas, após o termino da execução do algoritmo de coleta e remoção dos *links* duplicados, resultaram 84 páginas de notícia de violência contra LGBTQIA+ brasileiros distintas.

O conjunto dos links das 84 páginas resultantes do processo de coleta serviu então como entrada do algoritmo de extração de dados. Como resultado da extração, obteve-se então um conjunto de arquivos .XML contendo todas as informações extraídas das 84 páginas de notícias. Logo após, executou-se o algoritmo de integração de dados que integrou todos os dados presentes no conjunto de arquivos .XML e resultou no único arquivo .XML que constitui o banco de dados semi-estruturado gerado pela presente estratégia. Este arquivo contempla então todas os dados extraídos de todas as notícias unificados e serviu como entrada para a ferramenta de visualização de dados *Qlik Cloud*.

Dada a especificidade do algoritmo de integração, que é responsável por não armazenar dados duplicados provenientes dos mesmos casos de violência contra LGBTQIA+ no banco de dados semi-estruturado, este também foi experimentado de forma individual para avaliar sua precisão. Foram coletadas e selecionadas manualmente 20 notícias, de fontes distintas, que apresentavam informações sobre casos de violência em comum. Foram 7 conjuntos de notícias, onde cada conjunto apresenta informações sobre o mesmo caso de violência, que são compostos especificamente por: um conjunto contendo 5 notícias; um conjunto contendo 4 notícias; um conjunto contendo 3 notícias e quatro conjuntos contendo 2 notícias. Essas 20 notícias passaram pelo processo de extração de dados e integração de dados, com o objetivo único de avaliar a precisão do algoritmo de integração. Após a integração de dados ser realizada, avaliou-se cada um dos conjuntos e os resultados dos dados integrados buscando identificar a precisão do processo de integração.

Antes de submeter o banco de dados semi-estruturado, contendo dados sobre casos de violência contra a população LGBTQIA+ brasileira, para a ferramenta *Qlik Cloud*, houve um tratamento manual destes dados. Primeiramente, para cada *tag* notícia presente no banco, que contém todos os dados que foram extraídos da notícia, visitou-se o a página Web para validar os dados que foram devidamente extraídos. A validação ocorreu da seguinte forma: foram considerados acertos os dados extraídos que representavam exatamente a mesma informação presente no texto; foram considerados erros os dados que divergiam dos apresentados no texto ou que foram mencionados no texto, mas não foram extraídos. Caso algum dado não tenha sido mencionado no texto, mas o algoritmo de extração o tenha erroneamente identificado e extraído, também foi contabilizada a ocorrência de um erro. Visando ter uma visualização de dados corretos e reais, os dados identificados como incorretos foram removidos do banco de dados semi-estruturado para não gerar inconsistências nos dados estatísticos apresentados pela plataforma *Qlik Cloud*.

Por fim, na plataforma *Qlik Cloud*, ao fornecer o arquivo .XML que constitui o banco de dados semi-estruturado contendo dados sobre violência contra LGBTQIA+ como entrada, foi possível a geração e visualização de dados estatísticos referentes a violência contra a comunidade LGBTQIA+ brasileira.

4.3 Análise dos Resultados Obtidos

Nesta seção, são exibidos e analisados os resultados obtidos por meio da experimentação prática realizada, envolvendo a descrição experimental apresentada na Seção 4.2. A Subseção 4.3.1 apresenta os resultados e análises referentes ao processo de coleta semiautomática das páginas *Web* contendo notícias de violência contra LGBTQIA+ brasileiros. A Subseção 4.3.2, por sua vez, apresenta os resultados e análises da extração dos dados relevantes das páginas de notícia. Já a Subseção 4.3.3 descreve os resultados e análises do processo de integração dos dados. Por

fim, a Subseção 4.3.4 apresenta a visualização dos resultados obtidos por esta estratégia proposta, sendo estes os dados estatísticos referentes a violência contra a comunidade LGBTQIA+ obtidos a partir das páginas *Web* coletadas.

4.3.1 Coleta semiautomática de páginas *Web* contendo notícias sobre violências contra LGBTQIA+ brasileiros

Esta subseção apresenta e analisa os resultados obtidos a partir do processo de coleta semiautomática de páginas *Web* contendo notícias sobre violências contra LGBTQIA+ brasileiros, que representa o Passo 1 da arquitetura de funcionamento da presente estratégia (vide Figura 3.1).

Levando em consideração o algoritmo de coleta de páginas *Web* contendo notícias de violência contra LGBTQIA+ brasileiros, a Tabela 4.4 apresenta a quantidade total de páginas obtidas nos experimentos, a quantidade total de páginas selecionadas pelo usuário, a quantidade de páginas resultantes após a remoção das páginas duplicadas e a precisão do algoritmo com base nestes resultados obtidos.

Tabela 4.4 – Tabela com os resultados dos experimentos de coleta realizados

	Páginas obtidas	**Páginas selecionadas pelo usuário**	**Páginas resultantes após a remoção das duplicadas**	**Precisão**
Coleta	300	116	84	0,28

Fonte: elaborado pela autora

A precisão exibida na Tabela 4.4 foi calculada considerando a quantidade de páginas resultantes, após a remoção de possíveis páginas duplicadas selecionadas pelo usuário, em relação a quantidade total de páginas obtidas pelas execuções do algoritmo de coleta. A precisão de 0,28 para um algoritmo de coleta de páginas é relativamente baixa, significando que apenas cerca de 28% das páginas identificadas são realmente páginas de notícia de violência contra a comunidade LGBTQIA+ brasileira. Esta precisão sugere que o algoritmo está identificando muitas páginas como relevantes que, na verdade, não estão relacionadas à violência contra a comunidade LGBTQIA+ brasileira. Isto ocorreu devido ao conjunto de todos os termos de entrada não ter sido a combinação mais eficiente para a coleta de páginas *Web*, utilizando-se o buscador de páginas da *Google*, que contenham notícias de violência contra LGBTQIA+ brasileiros, já que os termos fornecidos ao buscador influenciam diretamente nos resultados.

4.3.2 Extração dos dados

Nesta subseção são apresentados e analisados os resultados obtidos a partir do processo extração de dados de páginas *Web* contendo notícias sobre violências contra LGBTQIA+ bra-

sileiros, que representa o Passo 2 da arquitetura de funcionamento da presente estratégia (vide Figura 3.1).

Notou-se, durante o processo de extração de dados, que algumas páginas de notícias bloqueavam o acesso ao conteúdo HTML de suas páginas. A biblioteca *BeautifulSoup* da linguagem python não conseguiu acessar e coletar o código HTML dessas páginas, consequentemente, seus textos não ficaram disponíveis para extração. Das 84 páginas de notícias fornecidas como entrada para o método de extração, 78 páginas permitiram a extração de seu conteúdo HTML. Das 78 páginas que obtiveram seus textos extraídos, apenas 64 eram realmente páginas de notícias sobre casos de violência contra a comunidade LGBTQIA+ brasileira. Alguns dos *links* selecionados como notícias de violência contra LGBTQIA+ brasileiros eram de casos internacionais ou eram páginas que continham apenas o vídeo da reportagem e pouco ou mesmo nenhum texto informativo. Somente a partir dos *links* destas páginas não era possível ter acesso a essas informações, por isso foram selecionadas. Estas páginas foram removidas no processo de validação dos dados extraídos.

Para validar se os dados extraídos estavam corretos, comparou-se os dados extraídos de cada página de notícias com o texto da respectiva página da qual foram extraídos. Conforme descrito na Subseção 3.1.2, para cada página de notícia, o processo de extração gera um arquivo XML contendo os dados extraídos da página e cada arquivo gerado contém o *link* para a notícia da qual os dados foram extraídos. Para cada arquivo, acessou-se a página de qual os dados foram extraídos e contabilizou-se os erros e acertos do processo de extração para os dados presentes na página. A Tabela 4.5 apresenta, para cada tipo de dado extraído, a quantidade total deste tipo de dado encontrado em notícias (já que nem todas as notícias possuem todos os tipos de dados considerados como relevantes), a quantidade de acertos na extração deste dado, a quantidade de erros na extração deste dado e a precisão do algoritmo baseada nestes resultados obtidos.

Tabela 4.5 – Tabela com os resultados dos experimentos de extração de dados realizados

Dado extraído	**Total**	**Acertos**	**Erros**	**Precisão**
Data	64	34	30	0,53
Nome	58	27	31	0,46
Idade	44	28	16	0,63
Orientação sexual	32	29	3	0,90
Identidade de gênero	35	35	0	1,00
Bairro	11	3	8	0,27
Cidade	53	17	36	0,32
Estado	55	17	38	0,30
Região	55	17	38	0,30
Assassinato	23	8	15	0,34
Agressão física	47	43	4	0,91
Agressão verbal	2	2	0	1,00
Privação de direitos	19	15	4	0,78

Fonte: elaborado pela autora

Tendo-se como base os diferentes métodos de extração de dados considerados como relevantes de páginas *Web* referentes a notícias sobre casos de violência contra LGBTQIA+ brasileiros, a partir Tabela 4.5, pode-se fazer análises distintas sobre a precisão do processo de extração de cada dado a partir das 64 páginas obtidas de notícia sobre violência contra LGBTQIA+ brasileiros:

- **Data em que a notícia foi publicada**: as 64 notícias apresentaram o dado referente a sua data de publicação. O processo de extração conseguiu extrair corretamente a data de 34 destas notícias, não conseguindo extrair a data de 30 notícias. Não ocorreu a extração de uma data indevida que não fosse a data de publicação da notícia; os 30 casos considerados como erros ocorreram quando a data estava presente na notícia, mas o algoritmo não conseguiu extraí-la. Isto deve-se ao fato destas páginas apresentarem uma estrutura HMTL da qual a biblioteca *BeautifulSoup* não conseguiu extrair a data como texto. Desta forma, o texto obtido a partir destas páginas de notícia não apresentava a data em seu conteúdo, impossibilitando assim a sua extração. A precisão de extração da data foi de 53% quando se considera as 64 notícias obtidas de violência contra LGBTQIA+ brasileiros.

- **Nome da vítima**: das 64 notícias, um total de 58 apresentaram o nome da vítima em seu conteúdo, destas quais foram extraídos corretamente os nomes das vítimas de 27 notícias. O algoritmo não conseguiu extrair o nome de 31 notícias, onde chegou a extrair nomes que não eram realmente os da vítima e até mesmo palavras que sequer eram nomes, mas começavam com iniciais maiúsculas. A extração de nomes que não eram os das vítimas ocorreu porque estes nomes foram citados mais vezes no texto do que o nome da própria vítima e o algoritmo considera o nome que mais ocorre no texto como sendo o da vítima. A identificação incorreta de palavras do texto como sendo o nome da vítima ocorreu devido ao rótulo "PER"do módulo *pt_core_news_lg* da biblioteca *SpaCy* tê-las reconhecido indevidamente como nomes. A precisão de extração do nome da vítima foi de 46% quando se considera as 58 notícias de violência contra LGBTQIA+ brasileiros que continham os nomes das vítimas em seus textos.

- **Idade da vítima**: das 64 notícias, um total de 44 possuíam a idade das vítimas mencionadas em seus textos. O algoritmo de extração conseguiu extrair devidamente a idade das vítimas de 28 notícias. Deixou de extrair a idade de 16 notícias, pelo fato de a idade ser apresentada, por exemplo, sem ser sucedida pela palavra anos, ou por não ter sido devidamente obtida no texto que é extraído do código HTML, já que o processo de extração do texto das páginas de notícia muitas vezes combina palavras sem espaços entre elas. A precisão da extração da idade da vítima foi de 63% levando em consideração as 44 notícias de violência contra LGBTQIA+ brasileiros que continham as idades das vítimas.

- **Orientação sexual da vítima**: das 64 notícias, um total de 32 apresentaram a orientação sexual da vítima em seus textos. Foi possível extrair a orientação sexual de 29 notícias e

deixou-se de extrair este dado de apenas 3 notícias. Isto ocorreu pelo falto de a orientação sexual ter sido citada de forma muito subjetiva nessas notícias como, por exemplo, "casal homoafetivo"e "casal de mulheres"; palavras que não correspondiam a nenhuma das palavras do conjunto que definem orientações sexuais a serem encontradas no texto. O algoritmo apresentou uma precisão de 90% ao extrair a orientação sexual das 32 notícias de violência contra LGBTQIA+ brasileiros que possuíam esse dado em seus textos.

- **Identidade de gênero da vítima**: das 64 notícias, 35 apresentaram a identidade de gênero da vítima em seus textos. O algoritmo conseguiu extrair a identidade de gênero de todas as 35 notícias, atingindo uma precisão de 100% quando se considera essas notícias. É um excelente indicativo que mostra que a estratégia de extração da identidade de gênero de notícias, apresentada neste trabalho, é eficaz.

- **Bairro em que a vítima sofreu violência**: das 64 notícias, um total de 11 apresentaram em seu texto o bairro em que a vítima sofreu violência. O algoritmo de extração conseguiu extrair devidamente o bairro de apenas 3 notícias, deixando de extrair o bairro de 8 notícias. Isto ocorreu devido a estas notícias não possuírem a palavra "bairro"precedendo o nome dos bairros citados no texto, palavra que era necessária para a extração do bairro definida nesta estratégia. A precisão de extração do bairro em que a vítima sofreu violência foi de 27%, tendo como base as 11 notícias obtidas de violência contra LGBTQIA+ que continham este dado.

- **Cidade em que a vítima sofreu violência**: das 64 notícias, 53 citaram a cidade em que a vítima sofreu violência em seus textos. O algoritmo conseguiu extrair devidamente a cidade de apenas 17 notícias, não conseguindo extrair a cidade do texto de 36 notícias. Aconteceu de outras cidades serem mais citadas no texto do que a cidade em que a vítima sofreu violência, levando a uma extração incorreta da cidade pelo algoritmo. O algoritmo também não conseguiu extrair devidamente os nomes da cidade de alguns textos de notícia devido ao rótulo "LOC"do módulo *pt_core_news_lg* da biblioteca *SpaCy* ter extraído do texto palavras que começavam com iniciais maiúsculas e não representavam nome de cidades. Quando o algoritmo tentava buscar essas palavras no arquivo que contém o nome de todas as cidades do Brasil, elas não eram encontradas, impossibilitando assim que realizasse a extração. Outro caso observado foi quando a cidade do Rio de Janeiro foi citada apenas como Rio no texto e o algoritmo não conseguiu extraí-la. Isto ocorre pelo fato de ele buscar pelo nome exato, extraído do texto, no arquivo que contém o nome de todas as cidade; se o nome estiver incompleto o algoritmo não consegue reconhecer como uma cidade, como foi o caso do Rio. A precisão do algoritmo de extração de cidades em que a vítima sofreu violência foi de 32%, levando em consideração as 53 notícias obtidas de violência contra LGBTQIA+ que continham este dado.

- **Estado e Região do Brasil em que a vítima sofreu violência**: das 64 notícias, um total de

55 notícias possibilitavam a extração do estado em que a vítima sofreu violência em seus textos. Nos moldes em que o algoritmo de extração foi implementado, a extração do estado em que a vítima sofreu violência dependia do nome da cidade e, por sua vez, a extração da região do Brasil em que a vítima sofreu violência dependia da extração correta do estado. O algoritmo extraiu devidamente o estado e região do Brasil de 17 notícias. Os casos em que a extração da cidade falhou, a extração do estado e região do Brasil consequentemente também falharam. O que notou-se, ao validar os dados coletados, foi que em alguns casos a cidade não era citada no texto, mas o nome do estado sim. Foram encontradas 2 notícias que contemplavam este cenário e, somadas ao total de notícias que continham o nome da cidade (53), têm-se um total de 55 notícias que possibilitavam a extração do estado em que a vítima sofreu violência em seus textos. Como a extração da cidade havia falhado para 36 notícias, somou-se também estes 2 casos em que não conseguiu-se extrair os estados do texto, quando estes foram diretamente citados, tendo um total de 38 notícias em que não conseguiu-se extrair os estados e regiões do Brasil em que as vítimas sofreram violência. A precisão do algoritmo de extração de estados e regiões do Brasil em que a vítima sofreu violência foi de 30%, levando em consideração as 55 notícias de violência contra LGBTQIA+ que possibilitavam a extração destes dados.

- **Tipos de violência sofridos pela vítima**: os tipos de violência sofridos pela vítima constituem em um conjunto de conceitos, definidos neste trabalho, sendo eles: assassinato, agressão física, agressão verbal e privação de direitos. Cada um desses conceitos de tipo de violência possui termos associados que são extraídos do texto, conforme visto na Subseção 3.1.8. Os resultados de erros obtidos neste processo de extração, que podem ser visualizados na Tabela 4.5, deram-se pelo fato das páginas de notícias apresentarem, em sua estrutura, manchetes de outras notícias que possuíam os termos que o algoritmo procura para classificar os tipos de violência. Ocorreu também de existirem termos que indicavam termos de violência que não constavam entre os buscados pelo algoritmo, o que impossibilitou o algoritmo de identificar e extrair alguns casos de violência.

  - **Assassinato**: das 64 notícias, um total de 23 notícias continham termos que são identificados pelo algoritmo de extração como pertencentes a casos de assassinatos. O algoritmo identificou corretamente esse tipo de violência em 8 notícias e acabou identificando erroneamente em 15 notícias. A precisão do algoritmo de extração para o tipo de violência assassinato foi de 34%, considerando-se as 23 notícias de violência contra LGBTQIA+ brasileiros que continham termos associados ao tipo de violência assassinato.
  - **Agressão física**: das 64 notícias, 47 notícias apresentaram em seus textos termos identificados pelo algoritmo de extração como sendo casos de agressão física. O algoritmo acabou deixando de reconhecer agressão física em 4 notícias, por não ter os termos que identificavam este tipo de violência em seus termos de busca para

extração. O algoritmo de extração apresentou uma precisão de 91%, tendo-se como base as 47 notícias de violência contra LGBTQIA+ brasileiros que continham termos associados ao tipo de violência agressão física.

- **Agressão verbal**: das 64 notícias, apenas 2 notícias apresentaram termos identificados como agressão verbal pelo algoritmo de extração. O algoritmo obteve uma precisão de 100% na identificação de termos associados ao tipo de violência agressão verbal, mas como foram poucas notícias que apresentaram esse tipo de violência, esse resultado de precisão não deve ser credibilizado. Uma base maior de notícias, com mais termos de violência verbal, proporcionaria um resultado de precisão mais realístico.
- **Privação de direitos**: das 64 notícias, um total de 19 notícias apresentaram termos que são identificados pelo algoritmo de extração como pertencentes a casos de privação de direitos. O algoritmo identificou e extraiu corretamente o tipo de violência privação de direitos de 15 notícias, não conseguindo realizar a extração para apenas 4 notícias. A precisão do algoritmo de extração para o tipo de violência privação de direitos foi de 78%, considerando-se as 19 notícias de violência contra LGBTQIA+ brasileiros que continham termos associados ao tipo de violência privação de direitos.

Como pode ser observado na Tabela 4.5 em conjunto com as análises dos resultados feitas nos tópicos anteriores, muitos dos processos de extração de dados definidos nesta estratégia apresentaram uma precisão considerável. A extração da data apresentou uma precisão de 53%, que poderia ter sido maior se as datas tivessem sido obtidas devidamente como texto a partir das páginas HTML das notícias. A extração da idade também apresentou uma precisão boa de 63%. A extração de orientação sexual, termos de agressão física e termos de privação de direitos apresentaram precisões excelentes, de 90%, 91% e 78% respectivamente. A precisão extração da identidade de gênero foi de 100% para as notícias coletadas, indicando que o algoritmo conseguiu extrair a identidade de gênero de todas as notícias que possuíam este dado. Já os processos de extração que não apresentaram uma precisão considerável, foi possível identificar os fatores que atrapalharam sua precisão, podendo estes serem corrigidos e melhorados em outras possíveis futuras versões da estratégia deste trabalho.

A primeira versão funcional da estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira, definida neste trabalho, apresentou uma precisão média de 59% em sua etapa de extração dos dados.

4.3.3 Integração dos dados

Esta subseção exibe e analisa os resultados obtidos a partir do processo de integração dos dados obtidos pelo processo de extração, processo este que representa o Passo 3 da arquitetura de funcionamento (vide Figura 3.1).

O processo de integração de dados, devido a sua especificidade de ser útil para cenários em que os mesmos casos de violência contra LGBTQIA+ brasileiros existem em fontes de notícias distintos, onde auxilia a não armazenar dados duplicados no banco de dados semi-estruturado, foi experimentado separadamente para avaliação de sua precisão, conforme descrito na Seção 4.2. A Tabela 4.6 apresenta os resultados obtidos dos experimentos do algoritmo de integração de dados.

Tabela 4.6 – Tabela com os resultados dos experimentos de integração de dados realizados

	Páginas obtidas como entrada	**Páginas integradas corretamente**	**Páginas que não foram integradas**	**Precisão**
Integração dos dados	20	9	11	0,45

Fonte: elaborado pela autora

Como pode ser visualizado na Tabela 4.6, das 20 páginas fornecidas como entrada ao experimento do algoritmo de integração, o algoritmo integrou corretamente 9 delas e deixou de integrar 11 páginas. As páginas que não foram devidamente integradas apresentaram uma divergência nos dados extraídos. Dado este cenário, o algoritmo de integração, ao comparar os dados extraídos, não as conseguiu identificar como sendo dados referentes a um mesmo caso de violência e, desta forma, considerou dados provenientes de notícias distintas.

A precisão do algoritmo de integração da presente estratégia foi de 45% considerando-se as 20 páginas de notícias de violência contra LGBTQIA+ fornecidas como entrada. Havendo melhorias no processo de extração de dados, é provável que estas desencadeiem um aumento na precisão do algoritmo de extração, já que este depende dos dados extraídos estarem corretos para poder compará-los e integrá-los.

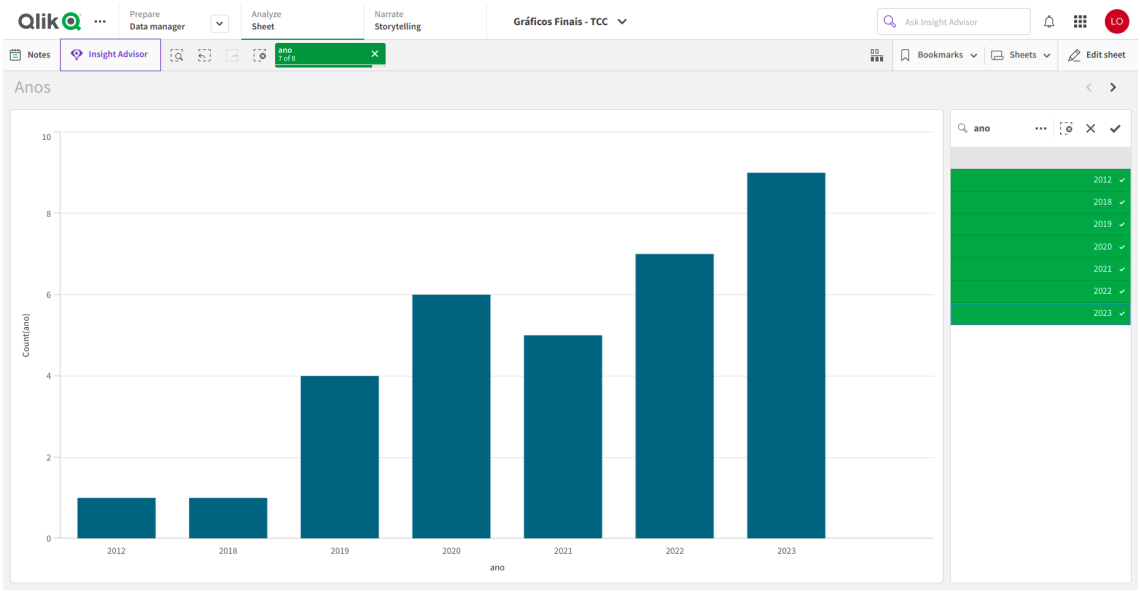
4.3.4 Visualização dos dados obtidos

Nesta subseção, são apresentados os gráficos obtidos a partir da primeira versão funcional da estratégia para geração semiautomática e dinâmica de dados estatísticos relativos a violência contra a população LGBTQIA+ brasileira definida neste trabalho. Esta subseção apresenta os resultados referentes ao Passo 4 da arquitetura de funcionamento (vide Figura 3.1).

Os dados incorretos obtidos pela primeira versão da presente estratégia foram removidos do banco de dados gerado, com a finalidade de gerar dados estatísticos que representem com fidelidade as informações presentes nas notícias coletadas. Isto foi realizado para que os gráficos gerados apresentem fielmente dados reais a partir das informações obtidas sobre violência contra a população LGBTQIA+ brasileira.O banco de dados semi-estruturado gerado, que consiste em um único arquivo XML, foi fornecido como entrada para ferramenta *Qlik Cloud*, onde foi possível gerar e visualizar dinamicamente gráficos contento dados estatísticos de violência contra a população LGBTQIA+ brasileira.

A Figura 4.1 apresenta o gráfico que exibe informações estatísticas sobre os índices de casos de violência contra LGBTQIA+ brasileiros em relação aos anos obtidos a partir das notícias coletadas.

Figura 4.1 – Índices de casos de violência contra LGBTQIA+ brasileiros em relação aos anos

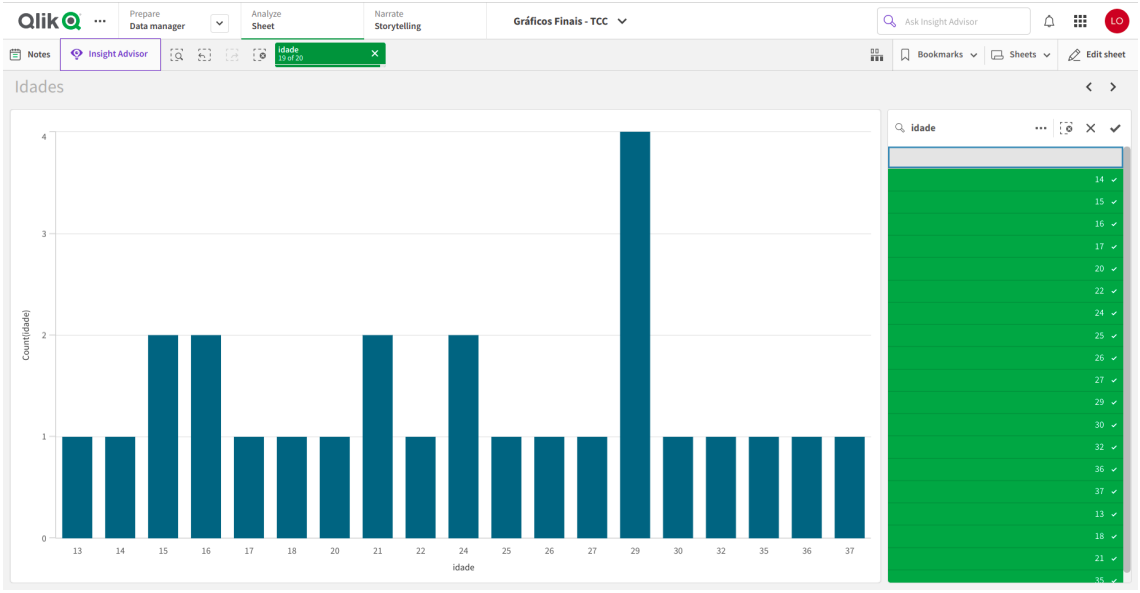


Fonte: elaborado pela autora

Como pode ser observado na Figura 4.1, é possível visualizar a contagem de casos de violência contra LGBTQIA+ ocorridos por anos, por meio dos dados obtidos através das notícias de violência contra essa população. O ano que mais apresentou casos foi 2023, com 9 casos de violência, devido ao fato de o buscador da *Google* priorizar o retorno de notícias que são mais recentes no processo de coleta.

A Figura 4.2 apresenta o gráfico que exibe informações estatísticas sobre a faixa etária das vítimas de casos de violência contra LGBTQIA+ brasileiros obtidos a partir das notícias coletadas.

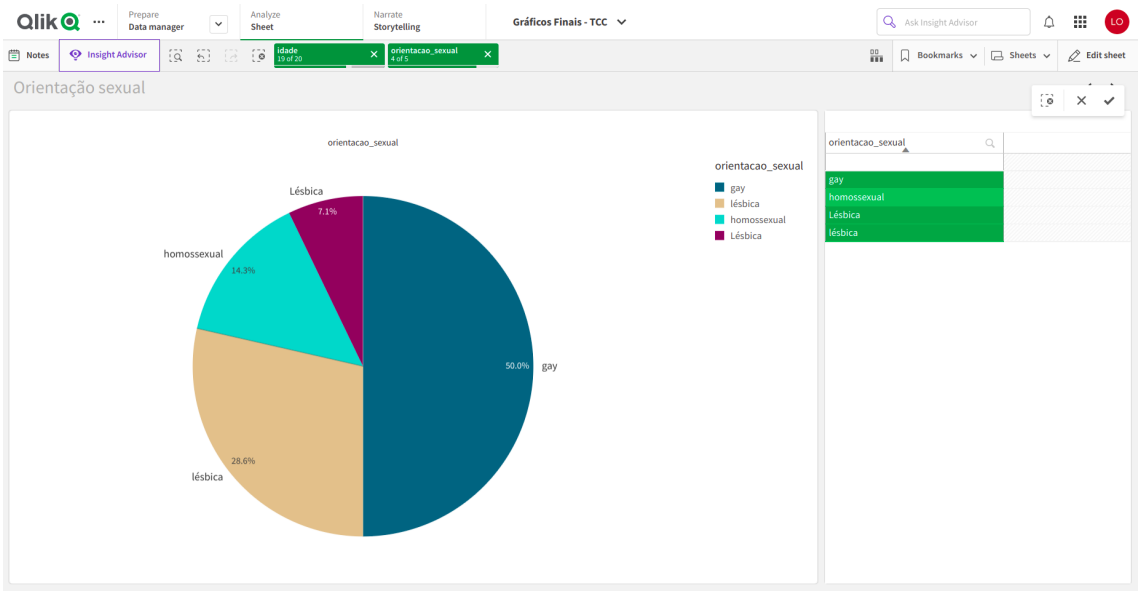
Figura 4.2 – Faixa etária das vítimas de casos de violência contra LGBTQIA+ brasileiros



Fonte: elaborado pela autora

Na Figura 4.3 pode-se visualizar o gráfico contendo os índices sobre a orientação sexual das vítimas de casos de violência contra LGBTQIA+ brasileiros obtidos a partir das notícias coletadas.

Figura 4.3 – Índices sobre a orientação sexual das vítimas de casos de violência contra LGBTQIA+ brasileiros



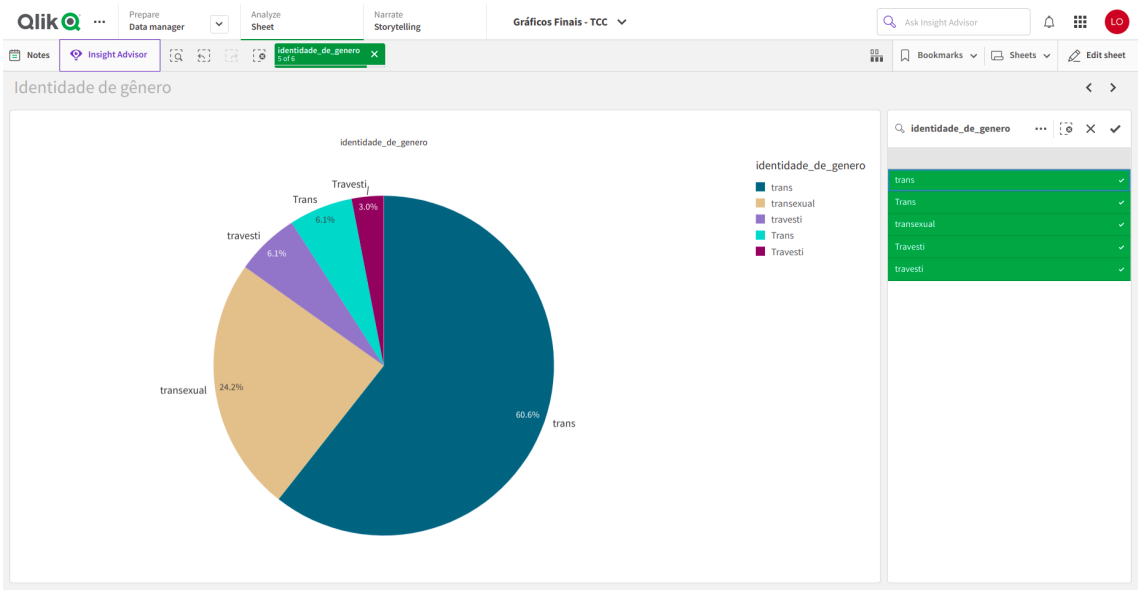
Fonte: elaborado pela autora

É possível visualizar no gráfico da Figura 4.3 que a plataforma *Qlik Cloud* faz distinção de termos iguais quando estes tem suas iniciais maiúsculas e minúsculas. Os dados referentes

aos termos "lésbicas"e "Lésbicas"referem-se a mesma orientação sexual e não deveriam ter sido apresentados como informações distintas no gráfico.

Já a Figura 4.4 apresenta o gráfico contendo os índices sobre a identidade de gênero das vítimas de casos de violência contra LGBTQIA+ brasileiros obtidos a partir das notícias coletadas.

Figura 4.4 – Índices sobre a identidade de gênero das vítimas de casos de violência contra LGBTQIA+ brasileiros

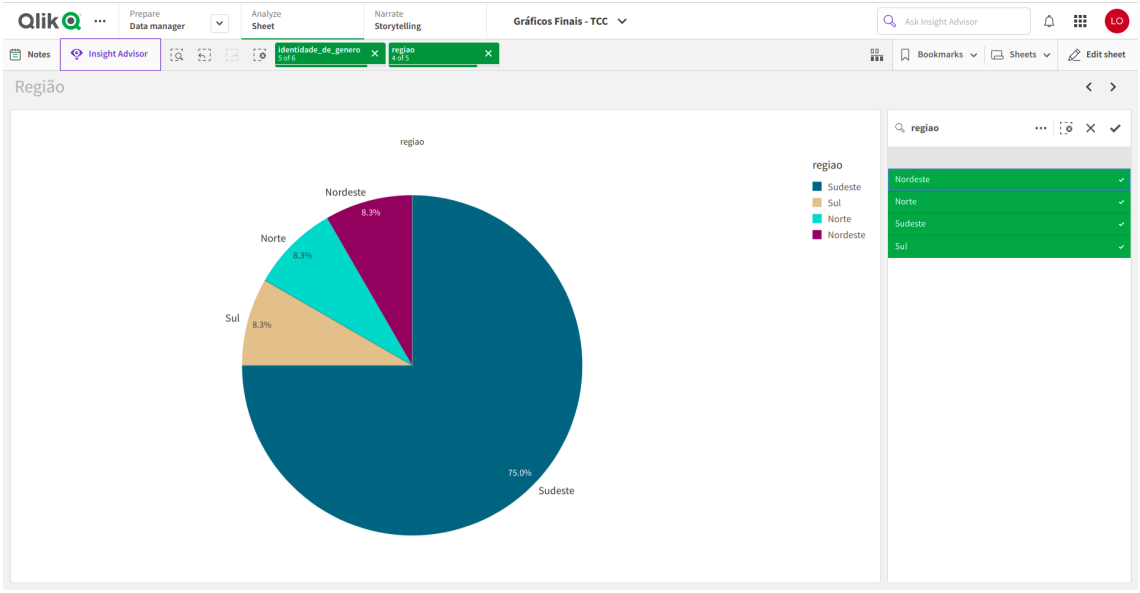


Fonte: elaborado pela autora

Observando-se o gráfico da Figura 4.4, percebe-se novamente a distinção de termos iguais quando estes possui suas iniciais maiúsculas e minúsculas. Os termos "trans"e "Trans"deveriam representar o mesmo dado no gráfico, assim como "Travesti"e "travesti". Nota-se também que os termos "trans"e "transexual"deveriam ter sido tratados de forma que houvesse uma associação que indicasse que estes termos referem-se a mesma identidade de gênero.

Por sua vez, a Figura 4.5 apresenta o gráfico contendo os índices de violência contra LGBTQIA+ nas regiões do Brasil obtidos a partir das notícias coletadas.

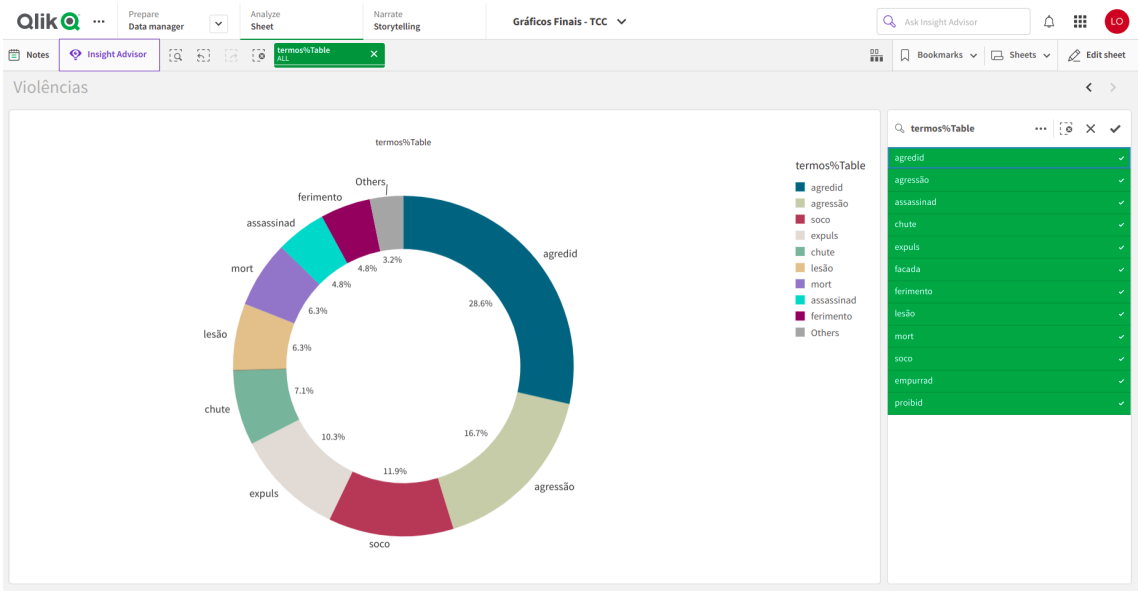
Figura 4.5 – Índices de violência contra LGBTQIA+ nas regiões do Brasil



Fonte: elaborado pela autora

Por fim, a Figura 4.6 apresenta o gráfico contendo os índices sobre os tipos de violências mais frequentes contra LGBTQIA+ brasileiros obtidos a partir das notícias coletadas.

Figura 4.6 – Índices sobre os tipos de violências mais frequentes contra LGBTQIA+ brasileiros



Fonte: elaborado pela autora

Os gráficos apresentados são exemplos de que a presente estratégia é capaz de gerar dinamicamente dados estatísticos, relativos à violência contra a população LGBTQIA+ brasileira, a partir de páginas *Web* de notícias. Outras associações entre os dados extraídos podem gerar outros

tipos de gráficos com diferentes informações estatísticas. Buscou-se nesta subseção apresentar os principais gráficos possíveis de se gerar a partir dos dados extraídos nesta primeira versão funcional da estratégia.

5 Considerações Finais

Este capítulo aborda as conclusões sobre o trabalho desenvolvido (vide Seção 5.1) e as perspectivas de trabalho futuro (vide Seção 5.2).

5.1 Conclusão

Como visto, este trabalho propôs e desenvolveu uma estratégia para geração semiautomática e dinâmica de dados estatísticos relativos à violência contra a população LGBTQIA+ brasileira.

Buscando-se avaliar uma primeira versão inicial da estratégia, como apresentado, foram realizados experimentos considerando-se os três módulos da estratégia: coleta de páginas *Web* de notícias sobre violência contra LGBTQIA+ brasileiros, extração dos dados destas notícias e integração dos dados. A precisão de 28% do algoritmo de coleta mostrou-se baixa, devido aos termos fornecidos como entrada não terem sido os mais eficientes. O algoritmo de extração apresentou uma boa precisão de 59% e os experimentos realizados indicaram também que alguns processos de extração podem ser melhorados, o que poderia levar a um aumento desta precisão. Por fim, o algoritmo de integração de dados apresentou uma precisão de 45% nos experimentos realizados, que provavelmente pode ser aumentada se houver melhorias no processo de extração de dados, já que o algoritmo de integração dos dados depende dos dados extraídos estarem corretos para poder compará-los e integrá-los. Com os dados gerados pela estratégia, foi possível visualizá-los na plataforma *Qlik Cloud* e obter informações estatísticas sobre índices de violências contra a população LGBTQIA+ brasileira por meio de gráficos.

Ademais, a versão inicial da presente estratégia mostrou que esta pode ser uma ferramenta útil, por meio dos dados extraídos e das visualizações de dados que podem ser criadas, para geração semiautomática e dinâmica de dados e evidências sobre a violência contra a população LGBTQIA+ brasileira.

5.2 Trabalhos Futuros

Nesta seção, são apresentadas algumas perspectivas de trabalho futuro. Desta forma, pretende-se: (1) melhorar o processo de coleta das páginas *Web* de notícias de violência contra a população LGBTQIA+ utilizando-se um coletor temático; (2) melhorar o processo de extração dos dados levando-se em consideração as melhorias identificadas nos experimentos da primeira versão funcional; (3) desenvolver e validar uma ferramenta para associação e geração de gráficos, a partir do banco de dados semi-estruturado gerado por esta estratégia.

Referências

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Answers to your questions: For a better understanding of sexual orientation and homosexuality.*: Disponível em: <http://www.apa.org/pi/about/newsletter/2008/04/brochureupdate.aspx>. acessado em 27 de abril de 2010. Washington, DC, 2008.

ASSIS, G. T. Uma abordagem baseada em gênero para coleta temática de páginas da *Web*. *UFMG*, v. 1, 2008.

BENEVIDES, B. *Dossiê assassinatos e violências contra travestis e transexuais brasileiras em 2021*: Disponível em: <https://antrabrasil.files.wordpress.com/2022/01/dossieantra2022-web.pdf>. acessado em 03 de fevereiro de 2023. Brasília, DF, 2022.

BERINATO, S. Good charts: The hbr guide to making smarter, more persuasive data visualizations. *Harvard Business Review Press*, s.d.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

BULGARELLI, L. et al. *LGBTIfobia no Brasil: barreiras para o reconhecimento institucional da criminalização*. São Paulo, SP, 2021. 13 p.

BUNEMAN, P. Semistructured data. *Sigmod International Symposium on Principles of Database Systems (PODS'97)*, Tucson, Arizona, USA, v. 16, p. 117–121, 1997.

BUTLER, J. *Problemas de gênero: feminismo e subversão da identidade*. [S.l.]: Civilização Brasileira, 2003. 59 p. ISBN 978-8520006115.

CAMPOS BRENO A.; CAMPOS, P. G. Análise de dados usando as ferramentas de business intelligence tableau e qlik sense. *Universidade de Pernambuco (UPE)*, s.d.

EBERENDU, A. C. Unstructured data: an overview of the data of big data. *International Journal of Computer Trends and Technology (IJCTT)*, v. 38, n. 1, 2016.

FORNARI, L. F. et al. Violência contra a mulher no início da pandemia da covid-19: o discurso das mídias digitais. *REME-Revista Mineira de Enfermagem*, v. 25, n. 1, 2021.

GOMES, M.; D'EMERY, R.; CYSNEIROS, G. Aapw: uma ferramenta para facilitar o aprendizado de programação web. p. 269–278, 2014.

GOMES, M. et al. A violência para com as pessoas lgbt: uma revisão narrativa da literatura. *Brazilian Journal of Health Review*, v. 4, n. 3, p. 13903–13924, 2021.

LIU, B. *Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data*. [S.l.]: Springer, 2011. v. 1. ISBN 978-3-642-19459-7.

LOPES, L. et al. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. *Reciis - Revista Eletrônica de Comunicação Informação & Inovação em Saúde*, v. 1, p. 13, 2009.

LOUSA, A.; PEDROSA, I.; BERNARDINO, J. Avaliação e análise de ferramentas business intelligence para visualização de dados evaluation and analysis of business intelligence data visualization tools. *Ieeexplore. Ieee. Org*, p. 19–22, 2019.

MACEDO, D. et al. Uma ferramenta para recomendação de visualização de dados governamentais abertos. *Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico*, p. 1–12, 2020.

MACÁRIO, C. G. d. N.; BALDO, S. M. O modelo relacional. *Instituto de Computação Unicamp*, Campinas, p. 1–15, 2005.

MARTINS, I. G. "justiça para dandara, Érika e para todas": a luta do movimento lgbt cearense por respostas estatais à barbárie transfóbica. *Universidade de Brasília*, p. 1–62, 2018.

MEDINA, A. P. R. Femini: jornalismo guiado por dados na construção de uma plataforma sobre violência contra mulher em são luís. *Universidade Federal do Maranhão*, p. 1–130, 2022.

MELLO, R. d. S. et al. Dados semi-estruturados. *Universidade Federal do Rio Grande do Sul (UFRGS)*, p. 39, s.d.

MINAYO, M. *Violência e saúde [online]*. Rio de Janeiro, RJ, 2006.

MONARD, M. C.; BARANAUSKAS, J. A. *Conceitos sobre aprendizado de máquina*. [S.l.]: Sistemas inteligentes-Fundamentos e aplicações, 2003. v. 1. 32 p.

MOREIRA, G. et al. *Manifestações de violência institucional no contexto da atenção em saúde às mulheres em situação de violência sexual. Saúde e Sociedade [online]*. [S.l.], 2020. v. 29, n. 1.

NASCIMENTO, H. A.; FERREIRA, C. B. Visualização de informações – uma abordagem prática. *XXIV JAI do XXV Congresso da Sociedade Brasileira de Computação*, 2005.

OLIVEIRA, C. o. *Estudo sobre a discriminação em função da orientação sexual e da identidade de gênero*. [S.l.]: Comissão para a Cidadania e a Igualdade de Género, 2010. 19 p. ISBN 978-972-597-326-4.

OLIVEIRA, J.; MOTT, L. Mortes violentas de lgbt+ no brasil: relatório 2021. *Grupo Gay da Bahia*, p. 1–78, 2022.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, p. 311–318, 2002.

PEREIRA, J. W. d. S.; CABRAL, J. N. A api do serviço filesender@ rnp. p. 17–23, 2023.

PINTO, M. B.; SACCOL, D. d. B. Um estudo sobre esquemas para documentos xml. *Centro Universitário Luterano de Palmas*, p. 1–10, 2003.

RIBEIRO, C. M. F. A. et al. Extração automática de conhecimento em documentos textuais: um estudo exploratório no domínio da sustentabilidade. *EmpíricaBR - Revista Brasileira de Gestão de Negócio e Tecnologia da Informação*, Rio Grande de Norte, v. 2, p. 14, 2015.

SANTAMARINA, Y. C. Balanceamento de carga em escalonamento de tarefas baseado em multi-commodity flow. Universidade Federal de Uberlândia, 2019.

SIMÕES, R. A. M. A importância dos dados estruturados, não estruturados e semiestruturados os desafios da sua utilização nas organizações brasileiras. *Universidade Federal Rural de Pernambuco*, v. 1, p. 45, 2022.

STODDER, D. Tdwi best practices report- visual analytics for making smarter decisions faster. *TDWI Research*, 2015.

TEIXEIRA, T. Infografia e jornalismo. conceitos análises e perspectivas. *EDUFBA*, 2010.

TORRES, G. M.; ZAINA, L. A.; ALMEIDA, T. A. de. Aprendizagem em redes socias: uma análise de dados do twitter. p. 87–90, 2012.

WANG, W. et al. Extracting 5w1h event semantic elements from chinese online news. *WAIM*, p. 644–655, 2021.

WIECZOREK, E. M. Recuperação de informação em documentos xml: Uma introdução. *Universidade Federal do Rio Grande do Sul*, p. 1–11, s.d.