UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DO CONHECIMENTO DE CIÊNCIAS EXATAS E ENGENHARIAS

LUCAS GERLACH NACHTIGALL

CRIAÇÃO E TRATAMENTO DE DATASET COM DADOS DE REDE SOCIAL PARA ESTUDOS SOBRE INTELIGÊNCIA ARTIFICIAL

CAXIAS DO SUL

2024

LUCAS GERLACH NACHTIGALL

CRIAÇÃO E TRATAMENTO DE DATASET COM DADOS DE REDE SOCIAL PARA ESTUDOS SOBRE INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul.

Orientador: Prof. Dra. Helena Graziottin Ribeiro

CAXIAS DO SUL

2024

LUCAS GERLACH NACHTIGALL

CRIAÇÃO E TRATAMENTO DE DATASET COM DADOS DE REDE SOCIAL PARA ESTUDOS SOBRE INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul.

Aprovado em 09/07/2021

BANCA EXAMINADORA

_____

Prof. Dra. Helena Graziottin Ribeiro
Universidade de Caxias do Sul - UCS

_____

Prof. Dra. Iraci Cristina da Silveira De Carli
Universidade de Caxias do Sul - UCS

_____

Prof. Dr. Odacir Deonisio Graciolli
Universidade de Caxias do Sul - UCS

RESUMO

A engenharia de dados envolve técnicas e ferramentas essenciais para coletar, organizar e transformar conjuntos de dados, garantindo sua qualidade e adequação para análises diversas. Este trabalho apresenta o desenvolvimento de um *dataset* sobre o tema da *Inteligência Artificial (IA)*, construído a partir de outros *datasets* abertos. Aplicando o método *Extract Transform Load (ETL)*, foram realizadas etapas de coleta, limpeza textual e padronização, utilizando ferramentas em Python para assegurar representatividade e qualidade dos dados. A validade do *dataset* foi testada em análises de sentimentos com os modelos *VADER* e *BERT*, comprovando sua versatilidade para diferentes contextos analíticos. Com isso, oferece uma base robusta para pesquisadores e estudantes desenvolverem novos estudos e investigações sobre *IA*, promovendo avanços na área de ciência de dados.

Palavras-chave: *ETL*, *Dataset*, Inteligência Artificial, Engenharia de Dados, Redes Sociais, Análise de Sentimentos, *VADER*, *BERT*.

LISTA DE FIGURAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

API *Application Programming Interface*
(Interface de Programação de Aplicações)

BD Banco de Dados

BERT *Bidirectional Encoder Representations from Transformers*
(Representações de Codificadores Bidimensionais de Transformadores)

BI *Business Intelligence*
(Inteligência de Negócios)

CSV *Comma-Separated Values*
(Valores Separados por Vírgula)

DW *Data Warehouse*
(Armazém de Dados)

ETL *Extract Transform Load*
(Extração, Transformação, Carregamento)

HTTP *Hypertext Transfer Protocol*
(Protocolo de Transferência de Hipertexto)

IA *Inteligência Artificial*

JSON *JavaScript Object Notation*
(Notação de Objetos JavaScript)

NLP *Natural Language Processing*
(Processamento de Linguagem Natural)

PBDA Portal Brasileiro de Dados Abertos

SGBD *Sistema de Gerenciamento de Banco de Dados*

SQL *Structured Query Language*
(Linguagem de Consulta Estruturada)

VADER *Valence Aware Dictionary and sEntiment Reasoner*
(Dicionário Sensível à Valência e Raciocinador de Sentimentos)

WWW *World Wide Web*
(Rede Mundial de Computadores)

SUMÁRIO

1 INTRODUÇÃO

A análise de dados é uma parte da ciência de dados, sendo uma área de excelência que integra estatística, computação, matemática e outras áreas. A combinação de técnicas dessas áreas é utilizada para extrair e analisar dados de formas diversas, nas mais variadas áreas de aplicação. Apesar de diversos trabalhos contendo seu foco na área de análise e exploração de dados existirem, há menos destaque à engenharia de dados, aos métodos de coleta e conversão aos quais os dados precisam ser submetidos, para que possam ser avaliados e depois analisados.

> Engenharia de dados é o desenvolvimento, implementação e manutenção de sistemas e processos que recebem dados brutos e produzem informações consistentes e de alta qualidade que suportam casos de uso *downstream*, como análise de dados e *machine learning*, A engenharia de dados é a combinação de segurança, gerenciamento de dados, DataOps, arquiterura de dados, orquestração e engenharia de software. Um engenheiro de dados gerencia o ciclo de vida da engenharia de dados, começando com a obtenção de dados dos sistemas de origem e terminando com o fornecimento de dados para casos de uso, como análise ou aprendizado de máquina (REIS; HOUSLEY, 2023).

Como dito por Reis e Housley (2023) a engenharia de dados abrange um conjunto de métodos, técnicas e ferramentas que, se bem utilizados, podem garantir a qualidade dos dados, que são a base para qualquer análise significativa. Sem informações organizadas e confiáveis, os métodos e processos analíticos não conseguem gerar resultados válidos.

Um artigo publicado pela revista *Forbes*[1] divulgou que a preparação de dados ocupava cerca de 80% do tempo gasto por cientistas de dados. O estudo foi realizado pela empresa *CrowdFlower* com um grupo de 80 cientistas (PRESS, 2016). Seis anos depois, um outro estudo, desta vez realizado pela empresa Anaconda e contando com 3493 indivíduos de 133 países, divulgou que ao contrário dos dados anteriores, os cientistas de dados gastam aproximadamente 54% do tempo na preparação e organização dos dados (ANACONDA, 2022).

Para tentar suprir esta necessidade, nos últimos anos surgiram eventos que possuem por objetivo servir como espaços de divulgação, além de promover a discussão de ideias e métodos de engenharia de dados para a construção de *datasets*, como por exemplo o Simpósio Brasileiro de Banco de Dados - *Dataset Showcase Workshop*[2], que atualmente se encontra em sua quarta edição, realizada em outubro de 2024.

Diante dessa situação, é importante contribuir com a descrição detalhada das diferentes tarefas e etapas que podem fazer parte do processo de preparação dos dados, além da produção de um conjunto de dados que pode ser utilizado em diferentes análises. Este trabalho visa desenvolver um *dataset* sobre a crescente utilização e propagação da IA, e conterá informações

[1] https://forbes.com.br/

[2] https://sites.google.com/view/sbbd-dsw/home

detalhadas sobre as ferramentas e técnicas utilizadas na criação do *dataset*, com o intuito de oferecer uma base para que profissionais e estudantes de computação possam utilizar os dados para suas pesquisas e análises.

Para a realização da coleta dos dados do atual trabalho, considerou-se inicialmente utilizar a rede social X[3] que é uma plataforma que gera uma grande quantidade de dados diariamente, com uma quantia de 564 milhões de usuários ativos mensais em todo o mundo, onde mais de 60% afirmam usar a rede para ficar informados sobre notícias (PEREIRA, 2023). A plataforma permitia, ainda enquanto se chamava Twitter, ser uma fonte de coleta de informações, em tempo real, sobre uma grande variedade de assuntos, como política, esportes, cultura e tecnologia, mantendo grande número de usuários ativos e garantindo boas condições para projetos de ciência de dados. Porém, a política de disponibilizar as API's de forma gratuita para coleta de dados sofreu mudanças com a venda do Twitter que se tornou X, e a coleta de dados passou a ser realizada com custo. Optou-se então por buscar *datasets* públicos com dados de Inteligência Artificial para que fossem integrados.

Com o aumento exponencial de dados gerados sobre Inteligência Artificial em diferentes plataformas, a criação de *datasets* organizados e confiáveis se torna essencial para apoiar pesquisas e estudos acadêmicos. Este trabalho busca preencher essa lacuna, oferecendo um *dataset* público e validado que pode ser usado para diversas aplicações analíticas, como análise de sentimentos e tendências tecnológicas.

1.0.1 Metodologia

Este trabalho utiliza o método ETL, que consiste em três etapas principais:

- **Extração:** Identificação e coleta de *datasets* públicos disponíveis sobre o tema de Inteligência Artificial, avaliando sua relevância e formato.
- **Transformação:** Limpeza, organização e integração dos dados, padronizando os formatos e eliminando inconsistências.
- **Carga:** Disponibilização do *dataset* final em plataformas públicas para acesso de pesquisadores e profissionais.

Adicionalmente, para validação do *dataset*, será realizado um estudo de caso utilizando técnicas de análise de sentimentos.

1.1 OBJETIVOS GERAL

A finalidade deste trabalho é desenvolver um *dataset* (Conjunto de dados) público, utilizando o método ETL, a partir de *datasets* públicos sobre a crescente utilização e propagação

[3] Rede Social, anteriormente conhecida como Twitter. Disponível em: https://www.X.com/

da IA.

1.1.1 Objetivos Específicos

Para atingir o objetivo geral, este trabalho propõe:

- Identificar e selecionar *datasets* públicos relevantes sobre Inteligência Artificial.
- Aplicar as etapas do método ETL (Extração, Transformação e Carga) para preparação dos dados.
- Analisar a qualidade dos dados tratados, garantindo consistência e confiabilidade.
- Demonstrar a aplicação do *dataset* desenvolvido por meio de um estudo de caso de análise de sentimentos.

1.2 ESTRUTURA DO TRABALHO

A estrutura do trabalho está organizada em capítulos que seguem uma abordagem lógica e progressiva. Cada capítulo se complementa para construir um entendimento completo sobre o desenvolvimento e utilização do *dataset*.

- No Capítulo 2, aborda-se os principais conceitos de *datasets* abertos, sua importância e como disponibilizá-los.
- No Capítulo 3, apresenta-se os principais conceitos de ETL, suas etapas e técnicas utilizadas para garantir a qualidade dos dados.
- No Capítulo 4, contextualiza-se sobre o uso das principais redes sociais, suas ferramentas para coleta de dados e sua importância no debate para o esclarecimento e desenvolvimento de ideias.
- No Capítulo 5, apresenta-se alguns trabalhos relacionados que abordam a criação e utilização de *datasets*, com dados provindos de redes sociais, em diferentes contextos.
- No Capítulo 6, são esclarecidos os principais tópicos da proposta para o desenvolvimento do *Dataset* final.
- No Capítulo 7, apresenta-se os processo de desenvolvimento do *dataset*, da integração dos datasets abertos, o tratamentos, a disponibilização, e um estudo de caso de sua utilização para análise de sentimentos para validação dos dados finais.
- No Capítulo 8, são exibidas as considerações finais sobre as finalidades atingidas do *dataset* proposto, além dos trabalhos futuros .

2 DATASETS

O uso de *datasets* (conjuntos de dados) tem se intensificado cada vez mais, especialmente em experimentos de *machine learning* e *data mining*. A importância de um bom conjunto de dados é que eles são a base para se obter êxito em qualquer análise, já que informações incoerentes podem resultar em conclusões imprecisas ou enganosas.

Há *datasets* abertos ou de acesso limitado. Os *datasets*, em especial a divulgação de informações em formato aberto, podem trazer vantagens importantes em várias áreas, possibilitando que pesquisadores, empresas e governos compartilhem e reutilizem dados nas mais diversas pesquisas e aplicações. Com a enorme quantidade de dados que são produzidos diariamente, a capacidade de disponibilizar, gerenciar e utilizar conjuntos de dados de forma adequada, tornou-se uma necessidade. Nas áreas de pesquisa, um conjunto de dados bem completo, organizado e estruturado é essencial para a realização de estudos e pesquisas, pois fornece a base a partir da qual modelos e inferências são construídos.

2.1 *DATASETS* ABERTOS

> O conceito de Dados Abertos aplica-se a todo dado publicado na Web disponível para que qualquer usuário possa utilizar, reutilizar e redistribuir esse dado sem qualquer restrição de patentes, propriedade intelectual ou outro mecanismo de controle, estando sujeito, no máximo, a atribuição de autoria.(MUNIZ, 2018)

Datasets abertos são conjuntos de dados disponibilizados publicamente sem restrições de uso. A disponibilização destes é uma prática cada vez mais comum em diversas áreas, incluindo órgãos governamentais, organizações e pesquisadores, com o objetivo de facilitar a transparência, a colaboração e a inovação. A utilização de dados abertos pode trazer benefícios significativos, tais como a melhoria de serviços públicos e o estímulo à criatividade e inovação. Junto disso, com o crescente número de sistemas de disponibilização de dados abertos, é possível que mais pesquisadores contribuam cada vez mais para a criação de soluções e aplicações inovadoras, que podem trazer benefícios para a sociedade como um todo.

A importância da definição e utilização de dados abertos, segundo *Open Knowledge Foundation (2024)*, está na interoperabilidade, que tem como seu significado a importante capacidade de diversos sistemas e organizações trabalharem juntos, a capacidade de diferentes componentes trabalhem juntos, ou combinar diferentes conjuntos de dados. A interoperabilidade , ou seja, a capacidade de dividir e de juntar componentes, é fundamental para a construção de sistemas grandes e complexos.

2.2 DISPONIBILIZAÇÃO DE *DATASETS*

A disponibilização de *datasets* é um processo que envolve a publicação de conjuntos de dados em um formato acessível e organizado, de modo que possam ser facilmente acessados, compreendidos e utilizados por outros usuários. A documentação adequada dos dados é essencial para garantir que os usuários possam entender e utilizar os dados de forma eficaz.

Nos dias atuais, há várias maneiras e formatos para disponibilizar dados abertos na web. Entre as plataformas de disponibilização de dados abertos mais conhecidas, estão *Kaggle*[1], Portal Brasileiro de Dados Abertos (PBDA) [2] e *Google Dataset Search*[3]. Esses são apenas alguns exemplos e existem muitos outros sites e plataformas disponíveis para compartilhar e encontrar conjuntos de dados.

Para a escolha de uma plataforma de compartilhamento de dados, depende-se de alguns aspectos, incluindo o público-alvo, o tipo de dados, o objetivo da publicação e as preferências pessoais. Cada uma das plataformas mencionadas anteriormente fornece instruções sobre como utilizá-las para o compartilhamento de *datasets*, sendo algumas mais restritivas (como o PBDA e o *Google Dataset Search*[3]).

Para o repositório *Kaggle*[1], por exemplo, sua principal orientação é compartilhar os dados em um formato acessível, visto que isso facilita o trabalho com os dados, independentemente da ferramenta utilizada, promovendo a acessibilidade e a reutilização por um público mais amplo. Existem algumas especificações técnicas como o limite de 200GB e 50 arquivos na raiz do *dataset*, além de dicas para adição de descrições sobre como documentar os dados, para facilitar a compreensão e o uso por outros usuários, através das tags e do campo descrição. Desde que o *dataset* siga as especificações técnicas citadas, só é necessário que existam os arquivos do *dataset*, um título, a informação sobre se tratar de um *dataset* privado ou público e qual a licença de publicação. Com isso disponível, pode se publicar um *dataset* (KAGGLE, 2024).

2.2.1 Formatos de Arquivo

Os formatos de arquivo são uma parte essencial da disponibilização de *datasets*, pois determinam como os dados são armazenados, organizados e acessados. Existem vários formatos de arquivo comuns para *datasets*, cada um com suas próprias características e vantagens, tornando-os mais adequado para diferentes cenários de armazenamento, análise e compartilhamento. Alguns dos formatos de arquivo mais utilizados, de acordo com kaggle (2024), incluem o CSV, JSON, *SQLite* e Arquivos, além das descrições de cada formato.

[1] https://www.kaggle.com/
[2] https://dados.gov.br/
[3] https://datasetsearch.research.google.com/

2.2.1.1 CSV

O CSV é um formato de arquivo mais simples, amplamente utilizado para armazenar dados tabulares. Nestes aquivos, os dados são organizados em linhas, onde cada valor é separado por uma vírgula ou ponto-e-vírgula. O CSV é um formato simples e de fácil leitura e que é diretamente transformado em tabelas (as vírgulas/ponto-e-vírgulas tornam-se as linhas, separando as colunas). Além disso, é um formato bastante flexível, permitindo a inclusão de diferentes tipos de dados, como números, textos e datas.

Um exemplo de *dataset* em CSV é mostrado na Figura 1, este sendo uma coletânea sobre *Pokemons*[4] e algumas de suas características.

Figura 1 – CSV Exemplo

abilities	against_bug	against_d...	against_dr...	against_el...	against_fa...	against_fi...	against_fire	against_fl...	against_g...
['Overgrow', 'Chlorophyll']	1	1	1	0.5	0.5	0.5	2	2	1
['Overgrow', 'Chlorophyll']	1	1	1	0.5	0.5	0.5	2	2	1
['Overgrow', 'Chlorophyll']	1	1	1	0.5	0.5	0.5	2	2	1
['Blaze', 'Solar Power']	0.5	1	1	1	0.5	1	0.5	1	1
['Blaze', 'Solar Power']	0.5	1	1	1	0.5	1	0.5	1	1
['Blaze', 'Solar Power']	0.25	1	1	2	0.5	0.5	0.5	1	1
['Torrent', 'Rain Dish']	1	1	1	2	1	1	0.5	1	1
['Torrent', 'Rain Dish']	1	1	1	2	1	1	0.5	1	1
['Torrent', 'Rain Dish']	1	1	1	2	1	1	0.5	1	1
['Shield Dust', 'Run Away']	1	1	1	1	1	0.5	2	2	1
['Shed Skin']	1	1	1	1	1	0.5	2	2	1
['Compoundeyes', 'Tinted Lens']	0.5	1	1	2	1	0.25	2	2	1

Fonte: Adaptado de Banik (2017)

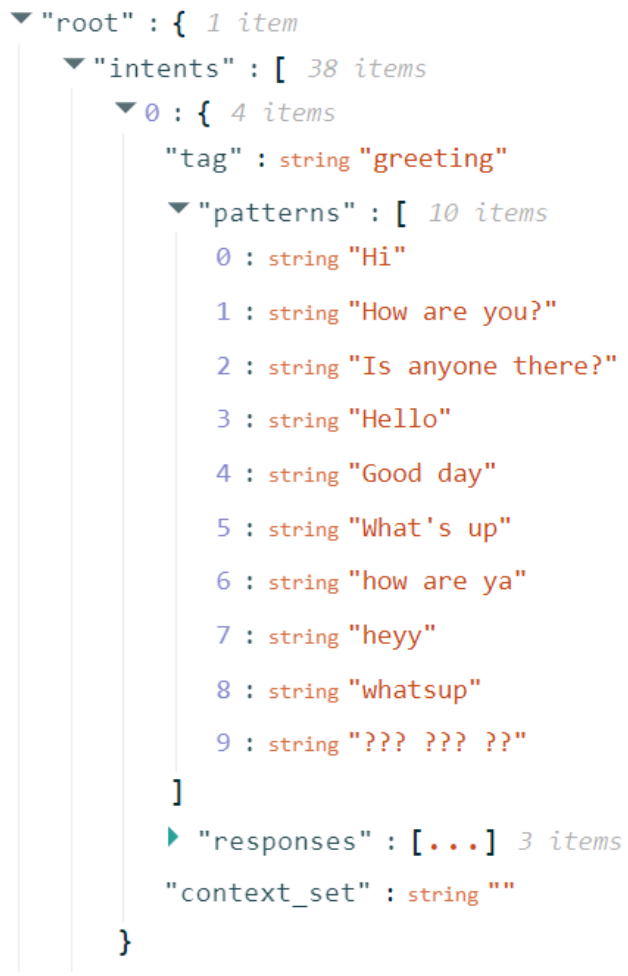
2.2.1.2 JSON

O JSON é comumente utilizado para armazenar e transmitir informações complexas e hierárquicas, ideal para representar estruturas semelhantes aos galhos de uma árvore. Nos arquivos JSON, os dados são estruturados em pares de chave-valor, com possibilidade de valores em *arrays* ou objetos aninhados, possibilitando uma representação abrangente e minuciosa dos dados.

Abaixo é apresentado um exemplo de *dataset* em JSON e mostrado na Figura 2, este sendo uma coletânea sobre possíveis falas de *chatbot*.

[4] Pokemon(abreviação de *Pocket Monsters*, Monstros de Bolso), são criaturas de uma franquia de mídia japonesa que inclui videogames, séries animadas, filmes e outros.

Figura 2 – JSON Exemplo



Fonte: Adaptado de Vaghani (2023)

2.2.1.3 *SQLite*

De acordo com a documentação do SQLite (2024), este é uma biblioteca em C que oferece um motor de banco de dados *Structured Query Language (SQL)* compacto, rápido e altamente confiável, que é amplamente utilizado em todo o mundo. A estrutura do *SQLite* é consistente e suportada em diversas plataformas. O código-fonte é de domínio público, o que significa que pode ser utilizado por qualquer pessoa sem restrições.

SQLite é um sistema de banco de dados leve e eficiente para gerenciar grandes volumes de dados, que permite a execução de comandos *SQL* sem a necessidade de um *Sistema de Gerenciamento de Banco de Dados (SGBD)*. Diferente do CSV, que é limitado a uma única tabela, o *SQLite* organiza os dados em múltiplas tabelas, assim permitindo uma melhor escalabilidade para grandes *datasets*. No *Kaggle*, cada tabela *SQLite* é visualizada separadamente na guia Dados, com suporte completo para metadados e métricas de colunas, similar aos arquivos CSV. Um ótimo exemplo de *dataset* no formato *SQLite* é o que se encontra na Figura 3.

Figura 3 – *SQLite* Exemplo

Table	Total Rows	Total Columns
Country	11	2
League	11	3
Match	25979	115
Player	11060	7
Player_Attributes	183978	42
Team	299	5
Team_Attributes	1458	25

Fonte: Adaptado de Mathien (2016)

2.2.1.4 Arquivos

Os arquivos não são *datasets* em si, mas quando se trata de grandes quantidades de dados, seja em áudio, imagens ou vídeos, eles podem ser comprimidos em formatos como ZIP e 7z, ideais para reduzir o espaço ocupado. Portanto, usar arquivos ZIP para conjuntos de dados extensos, sejam eles compostos por vários arquivos pequenos ou que estejam em subpastas, é uma ótima escolha de uso (KAGGLE, 2024).

Um exemplo de *dataset* em arquivos é mostrado na Figura 4, sendo uma coletânea de arquivos de imagens de raio-x do tórax de pacientes com diferentes tipos de pneumonia.

Figura 4 – Arquivos Exemplo

Fonte: Adaptado de Mooney (2018)

2.3 MONTAGEM DE *DATASETS*

A montagem de um *dataset* pode ser um procedimento complicado e extenso, porém é essencial em qualquer projeto de análise de dados ou *machine learning*. Um *dataset* bem organizado e estruturado assegura a precisão das análises e modelos desenvolvidos. No Capítulo 3, são abordadas as técnicas e estratégias sugeridas para a preparação de conjuntos de dados por meio do processo de ETL. A estrutura e conteúdo adequados de um *dataset* são em geral diretamente relacionados ao método ou modelo de *machine learning* que será utilizado. Então muitos dos *datasets* disponibilizados foram criados para serem utilizados em algum experimento específico, e podem ter que ser ajustados para alimentar outros experimentos ou métodos.

2.4 APLICAÇÃO DE *DATASET*

Dataset possuem uma ampla variedade de aplicações, desde tarefas simples, como classificação de texto, até análises mais avançadas em áreas como saúde, economia e análise de sentimentos. No contexto de análise de sentimentos, os *dataset* são utilizados para categorizar emoções em textos, enquanto em outras áreas podem ser aplicados para reconhecimento de padrões em dados biométricos, previsão de preços ou mesmo para treinar modelos de visão computacional em detecção de objetos (HUTTO; GILBERT, 2014; CRISTESCU *et al.*, 2023).

Os *dataset* desempenham um papel fundamental em análises de sentimentos e outras tarefas de *Natural Language Processing* (NLP). Eles fornecem dados estruturados necessários para o treinamento, validação e teste de modelos, permitindo que algoritmos aprendam padrões específicos e generalizem para novos cenários (ALJEDANI *et al.*, 2022; DEVLIN *et al.*, 2019).

No contexto de análise de sentimentos, *dataset* frequentemente contêm textos categorizados em polaridades, como positivo, negativo e neutro. Esses dados podem ser extraídos de redes sociais, avaliações de produtos e manchetes de notícias, entre outros. A qualidade e a representatividade de um *dataset* influenciam diretamente o desempenho dos modelos aplicados a ele, especialmente em abordagens supervisionadas de aprendizado de máquina (HUTTO; GILBERT, 2014; ALJEDANI *et al.*, 2022).

Além disso, mesmo dentro de uma única área, como análise de sentimentos, os modelos utilizados variam significativamente. Abordagens mais simples, como o VADER, priorizam rapidez e eficiência, enquanto arquiteturas avançadas, como o BERT, exploram a profundidade contextual e oferecem maior precisão em análises mais complexas (CRISTESCU *et al.*, 2023; DEVLIN *et al.*, 2019).

No contexto VADER é amplamente utilizado devido à sua simplicidade e eficiência em capturar polaridades emocionais em textos curtos, especialmente em redes sociais. O VADER combina um léxico robusto com heurísticas específicas para analisar a intensidade emocional de palavras e frases. Ele utiliza pontuações sentimentais atribuídas a termos individuais, ajustando

esses valores com base em fatores contextuais, como negações, modificadores de intensidade e uso de maiúsculas. Esse modelo alcança alta precisão em domínios como redes sociais e análises de serviços empresariais, sendo capaz de processar dados de forma eficiente, mesmo em grandes volumes de texto (HUTTO; GILBERT, 2014; YOUVAN, 2024).

Por outro lado, o BERT é um modelo de aprendizado profundo mais avançado, projetado para entender o contexto bidirecional das palavras em uma frase. Ele é baseado em uma arquitetura de transformadores e utiliza o aprendizado por transferência para realizar tarefas como análise de sentimentos, com um alto nível de precisão. Por exemplo, o BERT pode distinguir nuances contextuais, como ironias ou sarcasmos, que frequentemente escapam a modelos baseados em léxicos. Estudos destacam sua eficácia em tarefas complexas de NLP, especialmente ao ser ajustado com *dataset* específicos, demonstrando desempenho superior em relação a abordagens tradicionais, como VADER, embora com maior demanda computacional e tempo de processamento (DEVLIN *et al.*, 2019).

Essas abordagens ilustram o amplo espectro de ferramentas disponíveis para análise de sentimentos, desde soluções rápidas e especializadas, como VADER, até métodos mais avançados e adaptáveis, como BERT. Isso demonstra como os *dataset* podem ser moldados para atender a diferentes objetivos, destacando sua flexibilidade e importância na pesquisa e desenvolvimento.

3 ETL

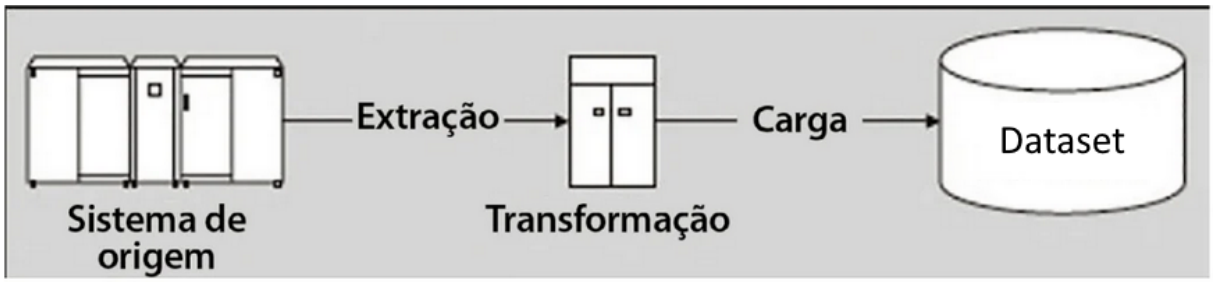
O processo de ETL visà a extração, transformação e carga de dados de uma ou mais fontes para uma ou mais bases de dados de destino, sendo este o processo mais crítico e demorado na construção de um *Data Warehouse* (DW), sendo este um repositório central de informações usadas para análise de dados(ABREU, 2008). Na sua origem, o processo de ETL era associado à construção e disponibilização de DW, mas atualmente é utilizado em diversos contextos, como a construção de *datasets* ou de Banco de Dados (BD), ou até para migrar dados de sistemas antigos para sistemas modernos.

O processo de ETL é uma estratégia bastante utilizada na construção de *datasets*. O processo consiste em três fases: (1) coleta de dados das fontes; (2) transformação para limpeza e formatação dos dados; e (3) carregamento dos dados no formato desejado (conforme ilustrado na Figura 5).

A construção de um *dataset* através de um processo de ETL adequado coleta informações de sistemas de origem, garante a qualidade e a coerência dos dados e os disponibiliza para utilização. Adicionalmente, o ETL é responsável por corrigir falhas e completar informações faltantes, assegurando a confiabilidade dos dados, protegendo o fluxo de dados transacionais, unindo dados de diversas fontes e preparando-os para serem utilizados por ferramentas de análise pelo usuário final. Em resumo, o ETL não apenas organiza os dados para a apresentação, mas também garante que sejam precisos, confiáveis e estejam prontos para atender às necessidades analíticas da organização (KIMBALL; CASERTA, 2011).

Devido a sua estrutura de funcionamento, a qualidade é aprimorada pelas soluções ETL ao realizar a limpeza dos dados (na fase de transformação de dados) antes de serem carregados em um repositório distinto (IBM, 2020).

Figura 5 – Plataforma de (ETL).



Fonte: Adaptado de VIDA *et al.* (2021)

3.1 EXTRAÇÃO (*EXTRACT*)

A extração de dados é a primeira etapa do processo ETL, onde os dados podem ser coletados de diversas fontes, alguns exemplos dados por VIDA *et al.* (2021) são: a partir de SGBDs, de ambientes em nuvem, híbridos e locais, de plataformas de armazenamento de dados, de outros *DW/datasets*, de documentos, de arquivos simples, de e-mails e de páginas da web através de APIs ou *web scraping*.

Nos dias atuais, a extração de dados de páginas da web tem se tornado uma prática comum, devido a grande quantidade de informações disponíveis na internet. A extração de dados de páginas da web pode ser feita de duas maneiras: através de APIs ou por *Web Scraping*. A extração de dados através de APIs é a maneira mais segura e eficiente de coletar dados de páginas da web, pois as APIs são projetadas para fornecer acesso a dados de forma estruturada e segura. Muitos sistemas e ferramentas fornecem APIs para que os interessados possam utilizá-las para consumir dados que estes fornecem.

De acordo com Broucke e Baesens (2018), a regra geral é procurar uma API primeiro e usá-la se possível, antes de começar a construir um *web scraper* para coletar os dados. No entanto, ainda existem várias razões pelas quais o *web scraping* pode ser preferível ao uso de uma API:

- O site do qual você deseja extrair dados não fornece uma API.
- A API fornecida é limitada: o que significa que você só pode acessá-la um certo número de vezes por segundo, por dia, etc.
- A API não fornece todos os dados que você deseja obter (enquanto o site fornece).
- A API fornecida não é gratuita (enquanto o site é).

Em todos esses casos, o uso de *Web Scraping* pode ser útil. É fato que se pode visualizar alguns dados em seu navegador da web, poderá acessá-los e recuperá-los por meio de um programa. Se você pode acessá-lo por meio de um programa, os dados podem ser armazenados, limpos e usados de qualquer maneira (BROUCKE; BAESENS, 2018).

3.1.1 API

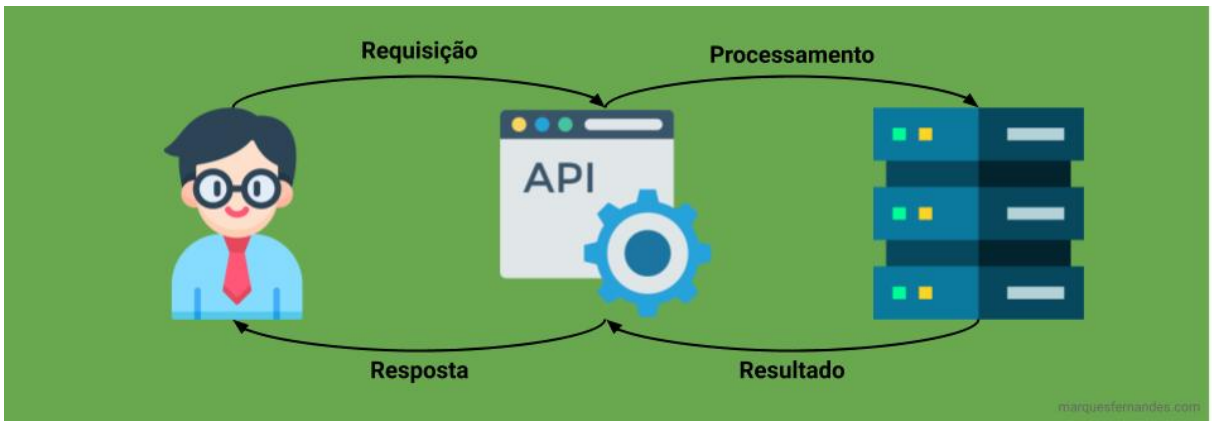
APIs consistem em ferramentas, definições e protocolos usados para desenvolver aplicações de software. APIs facilitam a conexão entre soluções e serviços, sem exigir conhecimento sobre a implementação desses elementos (REDHAT, 2018).

Normalmente, uma API é um software que facilita a comunicação de informações e recursos entre sistemas distintos. Ela especifica os procedimentos, a organização das informa-

ções e as diretrizes que os programadores devem obedecer ao se comunicar com um serviço específico. (REIS, 2023).

Na Figura 6, é mostrado como funciona a estrutura de uma API. O processo começa com a solicitação de dados de um cliente para o servidor. Em seguida, o servidor processa a solicitação e retorna os dados solicitados para o cliente. A comunicação entre o cliente e o servidor é feita através de uma API, que define as regras e procedimentos para a comunicação entre os dois sistemas (FERNANDES, 2020).

Figura 6 – Estrutura de uma API.



Fonte: Fernandes (2020)

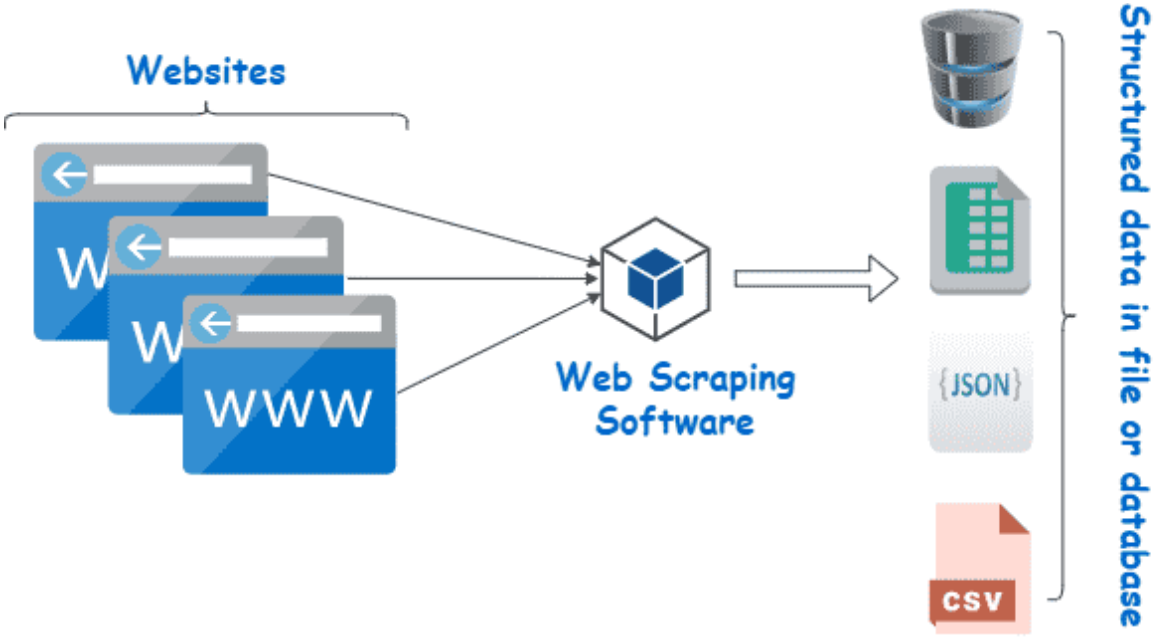
É comum que chaves de APIs sejam usadas para identificar e autenticar usuários de uma API. Elas são usadas para controlar o acesso a recursos e serviços, além de garantir que apenas um sistema autorizado possa acessar esses recursos. As chaves de APIs são frequentemente usadas para monitorar o desempenho e a segurança da API, na mesma medida em que auxiliam a proteger os recursos e serviços da API contra uso indevido (JACOBSON; BRAIL; WOODS, 2011).

3.1.2 Web Scraping

A técnica de *Web Scraping*, também chamada de extração ou coleta de dados da internet, consiste em extrair informações da *World Wide Web* (WWW) e armazená-las em um sistema de arquivos ou banco de dados para uso futuro. Normalmente, os dados da internet são obtidos usando o protocolo *Hypertext Transfer Protocol* (HTTP) ou através de um navegador web. Esse processo pode ser realizado manualmente pelo usuário ou de forma automatizada. Por causa do grande volume de dados variados sendo constantemente criados, o *Web Scraping* é bastante reconhecido como uma técnica eficaz e poderosa para a aquisição de dados (ZHAO, 2017).

Na Figura 7, é mostrado como funciona a estrutura de *Web Scraping*. O processo começa com a extração de dados de uma ou mais páginas da web. Em seguida, esses dados são armazenados em estruturas de dados apropriadas, estes podem ser salvos como arquivos (CSV, JSON, ou outros formatos) ou em bancos de dados.

Figura 7 – Estrutura de Web Scraping.



Fonte: kinsta (2023)

Web Scraping é amplamente utilizado em diversos campos da ciência da computação, especialmente nas principais tendências e novas áreas como: *Business Intelligence* (BI), IA, Ciência de Dados, *Big Data*, Computação em Nuvem e Segurança Cibernética (KHDER, 2021).

Por conta da necessidade de extração de dados por pesquisadores, algumas ferramentas e bibliotecas que utilizam *Web Scraping* surgem, como: *Octoparse*[1], *ParseHub*[2], *WebHarvy*[3], *Scrapy*[4], *Beautiful Soup*[5], *snscrape*[6], entre outras. Cada uma dessas ferramentas possui suas próprias características e funcionalidades, mas todas têm como objetivo a extração de dados de páginas da web, e podem ser utilizadas para coletar dados de redes sociais. Porém, casos como *Octoparse*, *ParseHub* e *WebHarvy* são ferramentas pagas, enquanto *Scrapy*, *Beautiful Soup* e *snscrape* são gratuitas.

3.2 TRANSFORMAÇÃO (*TRANSFORM*)

Na etapa de transformação, os dados não tratados são processados para serem alterados e unificados de acordo com a análise desejada. É fundamental que essas mudanças aconteçam na região intermediária para impedir impactos imediatos nos sistemas originais e diminuir a possibilidade de dados serem corrompidos. Ao efetuar essas alterações previamente, reduzimos o impacto no desempenho desses sistemas e diminuímos o risco de danos nos dados. Essa fase

[1] https://www.octoparse.com/
[2] https://www.parsehub.com/
[3] https://www.webharvy.com/
[4] https://scrapy.org/
[5] https://pypi.org/project/beautifulsoup4/
[6] https://pypi.org/project/snscrape/

é vista como a mais crucial no processo ETL, pois assegura aprimorar a qualidade dos dados e garantir sua chegada ao destino final de maneira adequada e preparada para ser utilizada (VIDA *et al.*, 2021).

Segundo Kimball e Ross (2013), a fim de garantir a precisão e qualidade de um conjunto de dados, alguns requisitos são essenciais, tais como:

- Informações precisas: Os dados devem conter valores e descrições precisos, ou seja, devem ser fiéis e verdadeiros.
- Evitar ambiguidades: Os dados devem ter valores e descrições com apenas um significado.
- Consistência: Os valores e explicações dos valores devem ser representados de forma consistente, utilizando apenas uma forma de notação para transmitir o seu significado, sem incluir mais do que uma para o mesmo dado.

Para garantir a qualidade dos dados, Hameed e Naumann (2020) sugere um conjunto variado de tarefas de preparação, tais como:

- Descoberta de dados: É o processo de analisar e coletar dados de diferentes fontes, seja para combinar padrões de dados, encontrar dados ausentes, localizar valores discrepantes, entre outros.
- Validação de dados: Compreende regras e restrições para inspecionar os dados. Por exemplo: corrigir, completar e verificar a qualidade dos dados.
- Estruturação de dados: Esta abrange tarefas para a criação, representação e estruturação de informações.
- Enriquecimento de dados: Serve para adicionar valor ou informações suplementares aos dados existentes de fontes separadas. Normalmente, envolve aprimorar os dados existentes com novos ou valores derivados de pesquisas sobre os dados, geração de chaves primárias e inserção de metadados.
- Filtragem de dados: Baseia-se em gerar um subconjunto dos dados em consideração, facilitando a inspeção manual e removendo linhas ou valores de dados irregulares. Exemplos incluem extração de partes de texto e manter ou excluir linhas filtradas.
- Limpeza de dados: Refere-se à remoção, adição ou substituição de valores de dados menos precisos ou imprecisos por valores mais adequados, precisos ou representativos. Exemplos típicos são deduplicação, preenchimento de valores ausentes e remoção de espaços em branco.

Existem casos onde se tem como objetivo a diminuição da complexidade e variabilidade de um texto, envolvendo a remoção de termos que não possuem nenhum tipo de relevância para o domínio analisado, assim como agrupar elementos idênticos (ou similares) que estejam apresentados de forma diferente. Para isso, existem processos de transformação de dados, como a remoção de *stop-words*, a manipulação de expressões regulares, a lematização e stemização, que são conhecidas como técnicas de NLP.

A aplicação de expressões regulares, lematização e stemização no processo de transformação dos dados visa reduzir a complexidade e a variabilidade textual, otimizando o conjunto de dados para análises subsequentes. As expressões regulares permitem identificar padrões e realizar manipulações textuais, como a remoção de caracteres irrelevantes e a padronização de formatos. A lematização e a stemização, por sua vez, possibilitam a unificação de termos semanticamente relacionados, eliminando redundâncias e simplificando a estrutura textual. Essas técnicas foram selecionadas por sua eficácia na minimização de ruídos e na melhoria da consistência e qualidade dos dados processados.

3.2.1 Integração de Dados

A integração de dados na etapa de transformação do ETL é crucial para garantir a qualidade e a consistência das informações que serão analisadas. Esse processo envolve consolidar dados provenientes de múltiplas fontes, estruturadas ou não estruturadas, de forma a preparar o conjunto de dados para análises subsequentes (IBM, 2020).

Durante a transformação, técnicas como agregação, deduplicação e mapeamento de dados são aplicadas. A agregação combina diferentes conjuntos de dados para gerar *insights* mais profundos, enquanto a deduplicação remove informações redundantes, aprimorando a precisão dos resultados. O mapeamento, por sua vez, associa os campos dos dados de origem aos formatos de destino, promovendo consistência (DOMO, 2024).

Outras práticas incluem a padronização de formatos, como unificar representações de datas ou normalizar unidades de medida, e o enriquecimento dos dados por meio de cálculos ou derivação de novas variáveis. Essas etapas garantem que os dados estejam alinhados com os objetivos analíticos e prontos para armazenamento em sistemas como DW(IBM, 2020; DOMO, 2024).

3.2.2 *Stop-Words*

As *stop-words* se refere-se às palavras que não possuem uma semântica significativa, mas são necessárias para a manutenção da gramática do texto que estão presentes, como preposições, artigos, conjunções, entre outros. Normalmente, *stop-words* são encontradas com maior frequência que as palavras que carregam conteúdo consigo, tornando-as removíveis, pois assim reduzem o tamanho do texto e melhoram a eficiência do processamento de dados (CARDONE;

MARINH; VAZ, 2012). Um exemplo de remoção de *stop-words* é mostrado abaixo:

Frase Original: A raposa marrom veloz pula sobre o cachorro preguiçoso.

Frase sem *Stop-Words*: raposa marrom veloz pula cachorro preguiçoso.

Stop-words removidas: A, sobre, o.

3.2.3 Expressões Regulares

Expressões Regulares tratam-se de expressões algébricas com símbolos de um dado alfabeto e caracteres operadores sobre um conjunto de cadeias de caracteres (strings) regulares, as quais possuem um tipo de notação para descrever um conjunto de caracteres. As expressões regulares representam essencialmente uma nova estruturação de padrões, com uma organização muito mais simples e direta. No caso da linguagem *Python*[7], por exemplo, é possível utilizar o módulo *RE*[8] para manipular expressões regulares (FERREIRA, 2019).

De acordo com Santos *et al.* (2022), há algumas classes de comandos de Expressões Regulares, como marcadores (que encontram determinados padrões), quantificadores (que permitem que se especifique a quantidade de vezes que tais padrões devem ser buscados) e âncoras (que servem para delimitar o início e fim de onde tais padrões devem ser encontrados).

3.2.4 Lematização e Stemização

A lematização e a stemização são técnicas de processamento de linguagem natural que visam reduzir as palavras à sua forma base, ou lema. A diferença entre as duas técnicas é que a lematização é mais precisa e complexa, enquanto a stemização é mais simples e rápida, por conta disso a stemização apresenta desvantagem em relação a lematização, já que a stemização pode gerar palavras inexistentes e criar uma maior distância entre palavras correlatas.

O processo de stemização refere-se a redução de palavras ao seu radical, ou tema. No processo de stemização, são removidos quaisquer afixos agregados ao radical de uma palavra. A stemização é feita a partir de uma busca, que se procura pela ocorrência de determinados radicais que podem ocorrer em cada palavra. Por seu jeito de funcionar, acaba reduzindo sua legibilidade, generalizando demais as palavras (SANTOS *et al.*, 2022).

Já a lematização se baseia na premissa de reduzir qualquer inflexão ou derivação de uma palavra para sua forma base, ou lema, que é a forma canônica de uma palavra. O vocabulário se torna simplificado, facilitando a capacidade preditiva dos modelos de aprendizado de máquina a serem aplicados. Um exemplo seria as variações do verbo estar (como por exemplo “estou”,

[7] https://www.python.org/
[8] https://docs.python.org/3/library/re.html

“estaremos”, “estará”), que seriam substituídas pela forma canônica do verbo: estar (SANTOS *et al.*, 2022).

3.3 CARREGAMENTO (*LOAD*)

No estágio final do processo ETL, os dados transformados são transferidos do local de preparação para o carregamento. Existem duas principais abordagens de carregamento: Carregamento Completo, onde todos os dados são carregados de uma vez como novos registros, o que é útil para manter um *dataset* completo, mas pode resultar em crescimento rápido e difícil de gerenciar; e Carregamento Incremental, que carrega apenas dados novos ou modificados ao comparar com registros existentes, tornando essa abordagem mais eficaz ao reduzir a duplicação e simplificar a gestão de grupos menores de dados (VIDA *et al.*, 2021).

4 REDES SOCIAIS

Segundo Souza (2010), uma rede social é definida como um serviço *Web* que permite que indivíduos: (1) construam perfis públicos ou semi-públicos dentro de um sistema, (2) articulem uma lista de outros usuários com os quais compartilham conexões e (3) visualizem e percorram suas listas de conexões e outras listas feitas por outros usuários no sistema, e não como um grupo de pessoas que interagem primariamente através de qualquer mídia de comunicação, já que este definiria a existência de redes sociais desde a criação da Internet.

Seguindo a definição de Souza (2010), existem diversas redes sociais que variam de acordo com suas finalidades. Entre as possíveis variações, existem:

- Redes profissionais. Exemplo: *LinkedIn*[1].
- Redes de amizade. Exemplo: *Facebook*[2].
- Redes voltadas para o compartilhamento de algum tipo específico de conteúdo. Dentre estas, são listadas como principais as redes de:
  - Blogs. Exemplo: *Reddit*[3].
  - Microblogs. Exemplo: *X*[4].
  - Imagens. Exemplo: *Pinterest*[5].
  - Vídeos. Exemplos: *Youtube*[6] e *TikTok*[7].

Em janeiro de 2024, o número de usuários de redes sociais em todo o mundo ultrapassou 5 bilhões, o que representa mais de 60% da população mundial, segundo KEMP (2024b). No Brasil, esse número chegou em 144 milhões de usuários, o que representa 66% da população brasileira, segundo KEMP (2024a). Com o crescimento do uso de redes sociais, estas se tornaram uma fonte valiosa de dados. Elas fornecem uma grande quantidade de informações sobre o comportamento humano, incluindo opiniões e tendências, que podem ser usadas para entender melhor a sociedade e prever eventos futuros. Estes são possíveis graças a disponibilização que tais redes fazem com seus dados, através de dados públicos, APIs ou iniciativas de dados abertos. Por causa disso, é possível coletar, analisar e interpretar esses dados para obter estudos para dentro e fora da rede (SANTOS *et al.*, 2022).

[1] https://www.linkedin.com/
[2] https://www.facebook.com/
[3] https://www.reddit.com/
[4] https://www.X.com/
[5] https://br.pinterest.com/
[6] https://www.youtube.com/
[7] https://www.tiktok.com/

Por meio destas redes sociais e dos dados que elas possuem, é possível analisar o comportamento humano sobre as sociedades urbanas em grande escala, contribuindo para diversas áreas de estudo, tais como a sociologia, psicologia, política, jornalismo, linguística, urbanismo, entre outros. Isso acaba por tornar as redes sociais em uma fonte valiosa de dados para pesquisas acadêmicas e científicas (SILVA *et al.*, 2019).

4.1 *REDDIT*

Nos últimos anos, sites de notícias sociais tornaram-se mais comuns. Estes sites são plataformas nas quais os usuários geram ou submetem links de conteúdos. As submissões são votadas e classificadas de acordo com seus votos totais, os usuários comentam sobre o conteúdo submetido e os comentários são votados e classificados de acordo com seu total de votos. Essas plataformas fornecem um tipo de Web-democracia que está aberta a todos os interessados. O maior exemplo deste tipo de plataforma é o *Reddit*[8] (WENINGER; ZHU; HAN, 2013).

Reddit é uma rede social que se desenvolve pela e em torno da própria comunidade, com estruturas auto-organizáveis, onde os próprios membros podem ser moderadores de comunidades, sendo responsáveis por criar e executar as regras de forma autônoma. Estas estruturas são chamadas de *subreddits*, que são comunidades dentro do *Reddit*, onde os usuários podem compartilhar conteúdo, fazer perguntas, discutir sobre um determinado assunto, entre outros. Cada *subreddit* tem suas próprias regras e moderadores, que são responsáveis por garantir que as regras sejam seguidas, garantindo uma forma de interação diretamente atrelada à cultura de seus membros (PROFERES *et al.*, 2021).

Existem atualmente mais de 100 mil comunidades ativas, onde os próprios usuários administram e moderam, os Administradores do *Reddit* raramente interferem em *subreddits*, a menos que haja violações das regras do site. Dentro dos *subreddits*, os usuários podem postar *links*, imagens, vídeos, textos, perguntas, etc. Os usuários podem votar positivamente ou negativamente em postagens e comentários, o que influencia a visibilidade do conteúdo. Os usuários também podem comentar nas postagens e responder aos comentários de outros usuários. Junto a estas opções de interação, os usuários possuem um sistema de *Karma*, que é uma pontuação que reflete a quantidade de votos positivos e negativos que um usuário recebeu em suas postagem e comentários. O *Karma* é uma forma de reputação na plataforma, que pode ser usado para avaliar a credibilidade de um usuário. Este sistema garante que usuários que possuem muito *Karma* possam contribuir com mais frequência (WENINGER; ZHU; HAN, 2013).

O *Reddit* fornece uma API, que permite acessar qualquer dado disponível publicamente na plataforma, incluindo postagens, comentários, perfis, comunidades e suas respectivas metainformações. Sendo possível recuperar dados históricos completos e em tempo real, além de ser possível coletar os dados completos das postagens, dados como comentários, votos, dados

[8] https://www.reddit.com/

do autor e os metadados destes. A API do *Reddit* é uma das mais completas e flexíveis, permitindo a coleta de dados de forma eficiente e eficaz, sendo uma fonte valiosa de dados para pesquisas (SANTOS *et al.*, 2022).

Apesar das grandes liberdades que o *Reddit* oferece, o trabalho com dados apresenta diversos desafios. Em razão dos *subreddits* serem geridos pelos próprios usuários, por exemplo, cada comunidade possui suas próprias regras e práticas de moderação e junto a isso a disparidade no tamanho das comunidades torna mais complexa a comparação entre as mesmas, podendo levar a resultados enviesados. Outro desafio é a alta anonimidade dos usuários, que permite que usuários mal-intencionados se comportem de forma antiética em comunidades que tenham regras mais brandas, o que pode levar a coleta de dados enviesados ou até mesmo a coleta de dados falsos (PROFERES *et al.*, 2021).

4.2 *FACEBOOK*

Facebook[9] é uma rede social criada em 2004 e até este momento já juntou mais de 111 milhões de usuário no Brasil, sendo equivalente a mais de 50% da população brasileira (KEMP, 2024a). Em 2021, o *Facebook* deixou de ser reconhecido por este nome e passou a se chamar *Meta*, refletindo uma nova direção para a companhia e a aposta no metaverso. O conglomerado é dono do *Facebook*, *Instagram*[10], *WhatsApp*[11] e *Messenger*[12] (SESSA, 2024). A plataforma foi criada com o objetivo de conectar pessoas e permitir que elas compartilhem informações, fotos, vídeos e mensagens com amigos e familiares (META, 2024).

A principal API que o *Facebook* possui para as aplicações lerem e gravarem dados na plataforma de rede social é a *Graph API*[13]. Grande parte das solicitações feitas à *Graph API* necessitam de um token de acesso, que pode ser gerado pela aplicação com a implementação do *Login* do *Facebook* (SESSA, 2024).

A disponibilidade dos dados depende de permissões concedidas pelo utilizador de forma a garantir a privacidade e segurança dos dados. Os desenvolvedores são obrigados a aderir às políticas de uso de dados do *Facebook*, que descrevem como os dados podem ser coletados, usados e compartilhados. Estas políticas enfatizam a importância de proteção e preservação da privacidade dos usuários (CANALTECH, 2024a).

Embora há a disponibilização de uma API pública, o acesso aos dados pode ser limitado devido a uma falta das permissões concedidas. Outra limitação é o fato dos dados que podem ser obtidos só poderem ser originados por contas famosas ou verificadas, o que inviabiliza estudos de comportamento, de postagem e de públicos menos representativos no *Facebook*. Embora há

[9] https://www.facebook.com/
[10] https://www.instagram.com/
[11] https://www.whatsapp.com/
[12] https://www.messenger.com/
[13] https://developers.facebook.com/docs/graph-api/

a alternativa de utilização de técnicas de *web scraping* para coletar os dados diretamente da plataforma, pode ser considerada uma violação dos termos de uso da plataforma, podendo levar a suspensão da conta ou até mesmo ações judiciais (SANTOS *et al.*, 2022).

4.3 *X*

Criado em 2003 por Jack Dorsey, Noah Glass, Biz Stone e Evan Williams, com nome de *Twitter*, esta é uma rede social gratuita que funciona como um serviço de *microblogging*. No início, a plataforma foi desenvolvida para ser utilizada por meio de mensagens de texto, chamados de *tweets*, com um limite de 140 caracteres. Em 2017, o limite foi aumentado para 280 caracteres. E em outubro de 2022 foi adquirida pelo empresário Elon Musk, que em abril de 2023 resolveu mudar o nome jurídico da plataforma de *Twitter Inc* para *X Corp*. Desde sua criação, a rede social se tornou um canal para que pessoas obtenham informações sobre as últimas notícias e para que pudessem expressar, em tempo real, suas opiniões sobre assuntos diversos (CANALTECH, 2024b).

Nesta rede social, existem algumas funcionalidades que servem para auxiliar na navegabilidade e interação entre usuário e assuntos, como por exemplo o uso de *hashtags*, que são palavras precedidas do símbolo de cerquilha (#), que quando selecionadas lhe redirecionam para ver outras postagens com a mesma *hashtag* usadas para demarcar um tópico de discussão. O símbolo arroba (@), sucedido do nome de usuário, serve para mencionar um usuário em uma postagem. Também é possível repostar, que significa compartilhar um *tweet* de outro usuário para todos os seguidores de quem está compartilhando. Outro recurso importante é a existência dos *Trending Topics*, que são tópicos ou *hashtags* determinados algoritmicamente como os mais populares no *X* naquele momento. Você pode escolher personalizar os *Trends* com base em sua localização e quem você segue. E por fim as *Threads*, que são uma série de postagens conectadas de uma pessoa. Você pode fornecer contexto adicional, uma atualização ou um ponto estendido conectando várias postagens juntas (X, 2024).

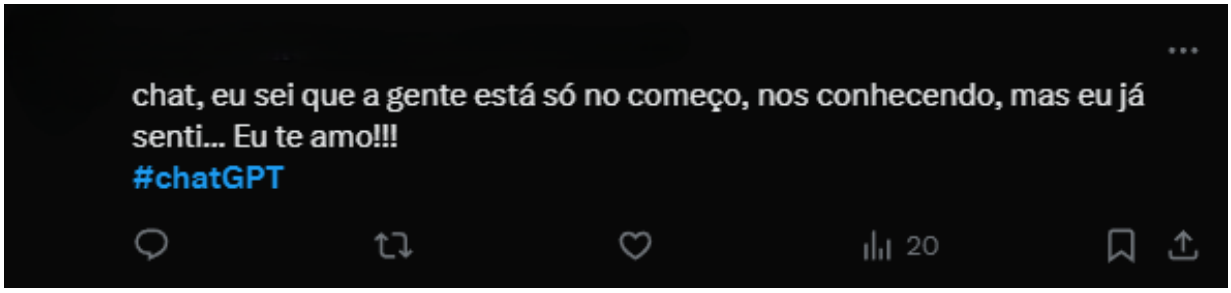
O *X*[14] é uma rede de informação composta por mensagens curtas (incluindo fotos, vídeos e links) de todo o mundo (X, 2024). Sendo usada por mais de 22 milhões de pessoas no Brasil (KEMP, 2024a), a plataforma cobre diferentes assuntos, como política, esportes, cultura e tecnologia. Essa variedade oferece uma visão ampla e inclusiva. Por sua estrutura de microblog e cultura de compartilhamento aberto de informações, se torna uma valiosa fonte de coleta de dados, essencial para pesquisas sociológicas e comportamentais. A maioria dos *tweets* são disponibilizados de forma pública, facilitando a obtenção de grandes quantidades de informações sem a necessidade de procedimentos complicados de autorização ou autenticação (SANTOS *et al.*, 2022).

Na Figura 8, se trata de um exemplo de *tweet* com reação positiva com o uso da IA

[14] https://www.X.com/

ChatGPT.

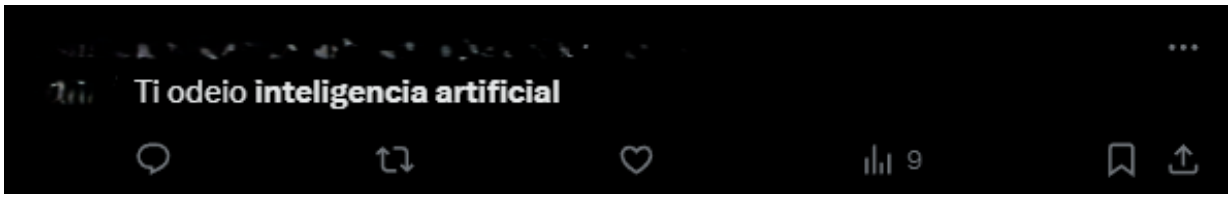
Figura 8 – Exemplo de *tweet* com uma reação positiva



Fonte: O Autor(2024).

A Figura 9, por outro lado, representa um exemplo de *tweet* com reação negativa a respeito de IA.

Figura 9 – Exemplo de Tweet com uma Reação Negativa



Fonte: O Autor(2024).

Os *tweets* citados anteriormente são exemplos de conteúdos que, quando obtidos em grande volume, podem ser relevantes para realizar uma análise de dados, como por exemplo uma análise de sentimentos ou análise de engajamento.

Após a compra da rede social pelo empresário Elon Musk, a plataforma passou por algumas mudanças. Uma das principais foi alteração dos planos grátis para utilização das APIs. Antes, com o plano grátis, era possível acessar duas versões da API (opção v2, mais moderna, e v1.1 clássica). Este plano permitia requisitar o limite de 2 milhões de *tweets* mensalmente e criar até 3 aplicações. Porém, após a mudança, todos os planos passaram a utilizar a API v2 e o plano grátis passou a permitir somente 1 aplicação e com um limite de 1500 postagens por mês, sem permitir requisição de dados. Ainda, foi criado um novo plano de assinatura, que possui benefícios que se aproxima dos benefícios do plano grátis anterior, sendo composto pela permissão de 3 aplicações e 1 milhão de requisições de *tweets* mensalmente. Este novo plano representa metade da capacidade de quando era gratuito, mas agora com o preço de 5 mil dólares por mês (X, 2024).

4.4 UTILIZAÇÃO DE DADOS DE REDES SOCIAIS

A utilização de dados de redes sociais, como *X*, *Facebook* e *Reddit*, tem sido fundamental para o avanço de técnicas de aprendizado de máquina, particularmente na análise de sen-

timentos. Essas plataformas oferecem um volume vasto de dados textuais, refletindo emoções e comportamentos em tempo real, que podem ser usados para explorar tendências e padrões sociais (ZHANG; QI, 2024; KAPUR; HARIKRISHNAN, 2022).

Além da análise de sentimentos, os dados provenientes de redes sociais têm demonstrado aplicabilidade em áreas diversas, como a previsão de tendências econômicas e o estudo do comportamento do consumidor. A literatura enfatiza que a riqueza desses dados está na sua representatividade de múltiplos contextos, abrangendo idiomas, culturas e perspectivas distintas, o que os torna essenciais para a modelagem preditiva e o aprendizado de máquina (KAPUR; HARIKRISHNAN, 2022; NIP; BERTHELIER, 2024). Contudo, o processamento dessas informações apresenta desafios significativos, como a remoção de ruídos, a mitigação de redundâncias e o enfrentamento de questões éticas, especialmente relacionadas à privacidade e representatividade.

Ferramentas que analisam a intensidade emocional, e modelos avançados de aprendizado por transferência, como o *BERT*, e abordagens lexicais, como VADER, destacam-se no tratamento desses dados, permitindo a extração de *insights* valiosos para a tomada de decisões baseadas em dados (KAPUR; HARIKRISHNAN, 2022).Conforme discutido na Seção 2.4, o VADER utiliza uma abordagem lexical para classificar o sentimento em categorias como positivo, negativo e neutro, enquanto o *BERT* aplica aprendizado por transferência para uma análise mais contextualizada e precisa.

Estudos recentes sugerem que a integração de técnicas de aprendizado profundo com a análise de redes sociais não apenas melhora a precisão dos modelos preditivos, mas também possibilita uma compreensão mais profunda das dinâmicas de interações digitais, criando oportunidades significativas para estratégias de negócios e avanços tecnológicos (DURGA *et al.*, 2023; NIP; BERTHELIER, 2024).

Em suma, a integração de dados de redes sociais com aprendizado de máquina oferece possibilidades significativas para o avanço científico, ao mesmo tempo que apresenta desafios técnicos e éticos que demandam atenção contínua. Estudos futuros devem explorar formas de mitigar essas limitações e maximizar o potencial analítico dessas plataformas.

5 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados alguns trabalhos relacionados que abordam a criação e utilização de *datasets* com dados provindos de redes sociais, em diferentes contextos. Nos estudos de ciência de dados, a extração, transformação e carga de dados (ETL) são processos fundamentais para a obtenção de informações relevantes e a construção de *datasets* de qualidade. A seguir, são apresentados alguns trabalhos que utilizaram a metodologia ETL para a criação de *datasets* a partir de dados coletados em redes sociais.

5.1 EXTRAÇÃO E AVALIAÇÃO DE UMA BASE DE DADOS SOBRE CRIMINALIDADE EM PORTUGUÊS A PARTIR DO *TWITTER*

Este trabalho propôs a extração e avaliação de uma base de dados sobre criminalidade a partir de publicações do *Twitter*. Foram feitas as coletas por meio da API do *Twitter* (MIRANDA *et al.*, 2023). Apesar de não comentarem sobre o processo ETL, a estrutura de extração, transformação e carregamento foi descrita durante o trabalho.

Durante a extração, foi utilizado um intervalo de tempo amplo, resgatando da rede social dados de 2010 a 2022, em todo território brasileiro. Durante a etapa de transformação, foram utilizadas duas bibliotecas *Python*[1]: a *OSMnx*[2] e a *Shapely*[3]. Estas foram utilizadas para filtrar somente os *tweets* da cidade de São Paulo a partir dos metadados de latitude e longitude presentes nos *tweets*. O carregamento foi feito em 3 arquivos CSV, que estão disponíveis no *Github*[4] com o nome de "repositorio_dados_sbcup"[5] (MIRANDA *et al.*, 2023).

A base de dados foi utilizada para a análise de crimes no município de São Paulo. Para a validação dos resultados, foram consultados os dados dos boletins de ocorrência no portal da transparência da cidade de São Paulo. Os resultados obtidos demonstraram a viabilidade da utilização de dados do *Twitter* para a análise de criminalidade. Embora tenham apresentados algumas limitações com a versão do Twitter para conseguir uma coleta precisa de *tweets* geo localizados, foi possível notar uma semelhança entre as bases oficiais e as informações de crimes coleta pelo Twitter (MIRANDA *et al.*, 2023).

O trabalho de (MIRANDA *et al.*, 2023) apresenta uma abordagem relevante para a extração e análise de dados textuais do Twitter, com foco na criminalidade, sendo uma referência importante para o presente estudo. Este estudo foi considerado na construção do *dataset*, especialmente no que diz respeito à utilização de metadados e ao potencial das redes sociais para

[1] https://www.python.org/
[2] https://osmnx.readthedocs.io/en/stable/
[3] https://shapely.readthedocs.io/en/stable/manual
[4] https://github.com/
[5] https://github.com/NESPEDUFV/repositorio_dados_sbcup

análises baseadas em grandes volumes de dados. Além disso, sua abordagem destaca a importância de integrar dados externos para validação, uma prática que pode ser explorada na expansões do *dataset* desenvolvido neste trabalho.

5.2 *DATA MINING OF SOCIAL MANIFESTATIONS IN TWITTER: AN ETL APPROACH FOCUSED ON SENTIMENT ANALYSIS*

Este estudo teve como objetivo analisar as reações dos usuários do *Twitter* a eventos sociais específicos, utilizando a metodologia ETL. O artigo "Bela, recatada e do lar", publicado pela revista Veja em 2016, foi escolhido como caso de estudo (YAGUI; MAIA, 2017).

O trabalho foi organizado em três fases principais: (1) coleta de informações, usando a API do *Twitter*, com as buscas em *tweets* com as palavras "bela, recatada e do lar"e a *hashtag* "#belarecatadaedolar"; (2) transformação dos dados, onde foi feita a filtragem de apenas mensagens em português e a remoção de caracteres especiais, pontuações e *stopwords*; e (3) carregamento dos dados, onde os dados foram carregados para um serviço de base de dados online chamado *Cloudant*, para que depois fossem salvos em uma máquina local e fosse feita a avaliação da polaridade dos sentimentos expressos nos *tweets* (YAGUI; MAIA, 2017). Apesar dos dados terem sidos carregados em um ambiente online, os mesmos não foram disponibilizados.

A combinação da metodologia ETL com o uso do algoritmo *Naive Bayes*[6] provou ser eficiente na avaliação de dados em redes sociais, fornecendo *insights* significativos sobre os comportamentos e opiniões dos usuários (YAGUI; MAIA, 2017).

Este trabalho apresenta uma abordagem relevante ao combinar a metodologia ETL com técnicas de aprendizado de máquina, como o *Naive Bayes*, para análise de sentimentos em redes sociais. Ele destaca a importância de processos estruturados de coleta e transformação de dados para a obtenção de *insights* significativos. Assim, contribui diretamente para a área ao demonstrar como métodos de mineração de dados podem ser aplicados em contextos específicos. As práticas empregadas nesse estudo servem de base para a construção de metodologias robustas e motivam o desenvolvimento de abordagens complementares no campo de análise de sentimentos, como as exploradas neste trabalho.

5.3 EXTRAÇÃO E TRATAMENTO DE DADOS DO *TWITTER* NO PERÍODO ELEITORAL DE 2022 PARA CRIAÇÃO DE UM *DATASET*

Este trabalho teve como objetivo a extração e tratamento de dados do *Twitter* no período eleitoral de 2022 para a criação de um *dataset*. A coleta de dados foi realizada por meio da API

[6] Naive Bayes é um algoritmo probabilístico de aprendizado de máquina baseado no Teorema de Bayes, utilizado principalmente na classificação de texto (MITCHELL, 2017).

do *Twitter*, com a busca de *tweets* relacionados às eleições presidenciais de 2022 no Brasil (LUNELLI, 2022).

A metodologia ETL foi utilizada para a extração, transformação e carga dos dados, que foram armazenados em 10 arquivos CSV e disponibilizados no site *kaggle*[7] com o nome de *Brazil Elections 2022 Twitter*[8] (LUNELLI, 2022).

Durante o processo ETL, o autor informou em detalhes todas as etapas para a construção do *dataset* final. A coleta de dados foi feita a partir de qualquer *tweet* que apresentasse algum dos termos seguintes: "votação", "votar", "bolsonaro", ou "lula". Para o tratamento, foi feita a limpeza dos *tweets*, assim removendo *emojis*, *links*, *stopwords* e duplicatas, juntamente com a formatação de horários UTC +00 para o UTC -03. O *dataset* foi criado para a análise de sentimentos dos usuários do *Twitter*[9] em relação aos candidatos à presidência. Os resultados obtidos demonstraram a viabilidade da utilização de dados do *Twitter* para a análise de sentimentos em períodos eleitorais, fornecendo *insights* significativos sobre as preferências dos eleitores (LUNELLI, 2022).

A construção detalhada e a disponibilização pública do *dataset* descrito por (LUNELLI, 2022) oferecem um exemplo claro de transparência e replicabilidade em projetos de ciência de dados. A limpeza textual e a remoção de duplicatas demonstram ser etapas indispensáveis para assegurar a qualidade e consistência dos dados. Ademais, a contextualização do período eleitoral enfatiza como *datasets* podem ser projetados para capturar eventos específicos, o que inspira a utilização estratégica de palavras-chave e *hashtags* na coleta de dados sobre Inteligência Artificial. Este trabalho destaca a importância da organização e documentação cuidadosa, que são fundamentais para que o *dataset* gerado possa ser reutilizado por outros pesquisadores.

5.4 CONCLUSÃO DO CAPÍTULO

Os trabalhos correlatos apresentam metodologias relevantes para a extração, tratamento e análise de dados do *Twitter*, especialmente em contextos sensíveis. O uso da metodologia ETL e o detalhamento das etapas de coleta e limpeza, como a remoção de duplicatas, *stopwords* e outros elementos irrelevantes, fornecem uma base sólida para a construção de *datasets* voltados à análise de sentimentos. A disponibilização dos dados no *kaggle* e os resultados alcançados demonstram a eficácia dessas abordagens e confirmam o potencial do *Twitter* como fonte de dados para estudos sociais e políticos. Este trabalho correlato contribui diretamente para este estudo ao oferecer técnicas e estratégias que são adaptadas e aplicadas na criação do presente *dataset*.

[7] https://www.kaggle.com/
[8] https://www.kaggle.com/datasets/eduardojoslunelli/brazil-elections-2022-twitter
[9] Atualmente X

6 PROPOSTA DE DATASET

O capítulo atual descreve a proposta inicial para o desenvolvimento de um conjunto de dados, a partir de *datasets* públicos, contendo dados obtidos através da rede social X[1]. O capítulo aborda os detalhes e as etapas envolvidas na criação do *dataset*, desde a coleta, tratamento, armazenamento e até a disponibilização dos dados. São abordadas as ferramentas e métodos utilizados para realizar esse processo.

Em um primeiro momento, é descrita a metodologia de extração de dados. A seguir, discute-se acerca das técnicas de tratamento de dados que são aplicadas para garantir a precisão e a integridade das informações, incluindo a remoção de duplicatas e a normalização dos dados. Posteriormente, mostra-se como os dados são agrupados em um *dataset* final, para assim serem disponibilizados em um formato acessível.

6.1 PROPOSTA

A finalidade deste estudo é desenvolver um conjunto de dados contendo informações sobre uso de IA. Este *dataset* é criado a partir de *datasets* públicos. O objetivo baseia-se em reunir informações relevantes e atualizadas sobre as tendências, opiniões e desenvolvimentos no campo de IA, que são posteriormente disponibilizadas ao público.

Os dados coletados são dos usuários da rede social *X*. Os dados são filtrados por assuntos e temas relacionados a IA, por nomes de algumas IAs tais como: *ChatGPT*[2], pertencente à empresa *OpenAI*, *Gemini*[3] da *Google* e à IA *Copilot* da *Microsoft*[4]. A escolha inicial pela rede social *X* (antigo Twitter) deve-se ao fato de ser uma plataforma com uma grande quantidade de dados gerados diariamente e por ser amplamente utilizada em discussões.

É feito um tratamento dos dados, com a remoção de elementos não essenciais do texto, com uma padronização do texto a partir da remoção de caracteres não alfabéticos, lematização e a remoção de duplicatas.

Ao final, o *dataset* é armazenado no formato CSV, por ser um formato simples e vastamente utilizado. Após os dados estarem devidamente organizados e armazenados, os mesmos são disponibilizados em um repositório público, para que possam ser usados em demais estudos, pesquisas e/ou análises.

[1] https://x.com/
[2] https://chatgpt.com/
[3] https://gemini.google.com/
[4] https://copilot.microsoft.com/

6.2 EXTRAÇÃO DE DADOS

Para a coleta dos dados, considera-se inicialmente o uso da API oficial da plataforma *X* (anteriormente conhecida como Twitter). No entanto, a utilização da API revela-se um desafio significativo devido às suas restrições de acesso e elevados custos associados ao número de requisições necessárias para coletar uma quantidade expressiva de dados. A plataforma apresenta diversas limitações que geram custos muito elevados, tornando essa abordagem inviável.

Além disso, técnicas de *web scraping*, que poderiam ser alternativas para a coleta dos dados, são descartadas devido às políticas restritivas da plataforma em relação à extração automatizada de dados, o que torna essa prática inviável.

Diante dessas dificuldades, opta-se por utilizar conjuntos de dados previamente coletados de redes sociais e disponibilizados publicamente na plataforma *Kaggle*[5]. Esses *datasets* atendem aos requisitos de conter *tweets* relacionados à Inteligência Artificial, com informações relevantes para a análise proposta, como menções a ferramentas como o *ChatGPT* e a IAs.

A utilização desses dados é perfeitamente válida dentro de um processo de ETL (Extração, Transformação e Carga), já que o objetivo da pesquisa é utilizar dados existentes para construir um novo conjunto de dados focado em análises relacionadas à IA.

6.3 TRATAMENTO DE DADOS

Após a coleta dos textos dos *tweets*, é necessário realizar a integração e tratamento dos dados. Nos textos, podem estar presentes links, *emojis*, *stopwords*, dentre outros caracteres, que precisam ser tratados ou removidos para a construção do conjunto de dados.

Para a realização do processo de limpeza, utiliza-se a biblioteca *pandas*[6]. O uso da biblioteca *pandas* é escolhido devido à sua robustez e versatilidade para manipulação de grandes volumes de dados, bem como pela ampla documentação disponível, o que facilita a reprodução do processo por outros pesquisadores. Os datasets a serem utilizados têm conteúdo semelhante (textos extraídos do Twitter), mas apresentam algumas diferenças nas estruturas que devem ser compatibilizadas. Em caso de haver *links* ou menções no texto, estes são removidos, pois não acrescentam valor semântico para a análise. *Hashtags* e *emojis*, por sua vez, não são mantidos na base do texto, mas sim em uma nova coluna para cada, pois podem acrescentar sentimentos e intenções.

Adota-se uma padronização com caracteres minúsculos e sem acentuação, com o intuito de unificar a forma de apresentação do texto e realizar a remoção de caracteres não alfabéticos, junto com uma lematização das palavras, para assim garantir uma estrutura mais limpa para análises futuras. Por fim, realiza-se a remoção de duplicatas, pois podem surgir dados em dupli-

[5] https://www.kaggle.com/
[6] https://pandas.pydata.org/

cata, que precisam ser identificados e tratados (removidos). Para isso, utiliza-se o ID do *tweet* como valor único, removendo assim as duplicatas.

6.4 VALIDAÇÃO DO DATASET: ANÁLISE DE SENTIMENTOS

Após a conclusão do processo de extração e tratamento dos dados, realiza-se uma análise de sentimentos como etapa de validação do conjunto de dados. O objetivo desta análise é verificar a qualidade e a consistência dos dados processados, bem como demonstrar sua usabilidade em contextos reais de aplicação.

A escolha da análise de sentimentos deve-se à sua relevância no campo de estudo de redes sociais e Inteligência Artificial. Este tipo de análise é amplamente utilizado para identificar tendências, opiniões e emoções relacionadas a temas específicos. Para este trabalho, realiza-se um estudo de caso, utilizando o *dataset* desenvolvido, para detectar sentimentos em menções às ferramentas de IA.

Essa abordagem garante que o *dataset* seja validado não apenas em termos de integridade técnica, mas também em sua aplicabilidade prática em diferentes cenários analíticos.

6.5 AGRUPAMENTO E DISPONIBILIZAÇÃO DOS DADOS

Após a obtenção do conjunto de dados, cria-se um arquivo CSV, contendo os dados referentes aos *tweets*. O formato CSV é escolhido para o armazenamento do *dataset* devido à sua simplicidade e ampla compatibilidade com diversas ferramentas analíticas, facilitando seu uso por pesquisadores e profissionais de diferentes áreas. Para isso, utilizam-se mais algumas funções da biblioteca *pandas*[7], que resultam em uma tabela contendo suas colunas conforme listado a seguir:

- **ID**: A chave primária, que contém o ID do *tweet*.
- **Date**: A hora e data do *tweet*.
- **Tweet**: O texto presente no *tweet*.
- **Hashtags**: Uma lista das *hashtags* utilizadas no *tweet*.
- **Retweets**: Quantidade de *retweets* do *tweet*.
- **Likes**: Quantidade de favoritos do *tweet*.
- **Emoji**: Uma lista com os emojis utilizados no *tweet*.

[7] https://pandas.pydata.org/

Após o tratamento e carregamento do conjunto de dados, ele é disponibilizado publicamente para análises posteriores. Inicialmente, o *dataset* é compartilhado no *Kaggle*[8]. A publicação inclui um título e dados relevantes para o *dataset* e *tags*, além de uma breve descrição; detalhes sobre a publicação do *dataset* estão no próximo capítulo 7.5.

[8] https://www.kaggle.com/

7 DESENVOLVIMENTO

Este capítulo apresenta o processo completo de desenvolvimento do *dataset* proposto no Capítulo 6. São detalhadas as etapas de coleta e formatação dos dados, a disponibilização do conjunto final, além da aplicação de análises que classificam os *tweets* em três categorias de sentimento: positivo, neutro e negativo.

7.1 COLETA E FORMATAÇÃO INICIAL

Para este trabalho, três *datasets* foram coletados manualmente através do *Kaggle*, utilizando os tópicos "*IA*", "*Twitter*", "*ChatGPT*", "*Copilot*" e "*Gemini*" como critérios de busca. As principais características de cada *dataset* estão descritas a seguir:

1. ***Dataset* 1: ChatGPT Launch Discussions on X (formerly Twitter)**
   Período: 30 de novembro de 2022 a 11 de fevereiro de 2023
   Autores: Waqas Aman e Aqdas Malik
   Descrição: Contém 948.116 valores únicos sobre *tweets* relacionados ao *ChatGPT* nos três primeiros meses após o lançamento. Os dados incluem texto dos *tweets* e metadados limitados, mas não possuem campo de `ID` e foram anonimizados para garantir privacidade.

2. ***Dataset* 2: 100K Tweets about ChatGPT**
   Período: 18 a 21 de março de 2023
   Autores: Anil
   Descrição: Reúne 92.559 valores únicos de *tweets* em inglês contendo a palavra "chatgpt". Inclui campos como `ID`, `Date`, `Username` (mascarado), `ReplyCount`, `RetweetCount`, `LikeCount` e `QuotesCount`.

3. ***Dataset* 3: Chatgpttweets**
   Período: 5 de dezembro de 2022 a 10 de junho de 2023
   Autores: Nora Abdo
   Descrição: Compreende 548.599 valores únicos, abrangendo colunas relacionadas ao perfil do usuário (`user_name`, `user_location`, `user_description`, `user_followers`, e `date`). Não possui coluna `ID`.

Para assegurar a consistência e a integridade dos dados durante a análise, foi realizada uma padronização das colunas em todos os *datasets*, essa padronização foi essencial para a unificação dos dados e facilitou o processamento subsequente, garantindo uma uniformidade na qualidade do conjunto final.

A unificação de *datasets* distintos apresentou uma série de desafios complexos, tanto na padronização dos dados quanto na garantia de sua integridade e consistência. Cada um dos *datasets* possuía uma estrutura de colunas específica e diferente dos demais, com campos variados em termos de nomenclatura e ordem, além de algumas colunas inexistentes, como o identificador único de cada tweet (`ID`). Esse cenário exigiu um tratamento específico para cada *dataset*, com etapas de transformação que permitiram converter os dados em uma estrutura uniforme. O processo de padronização foi essencial para garantir que os dados pudessem ser combinados de forma precisa e coerente.

7.2 FILTRAGEM DE *TWEETS* EM INGLÊS

Para assegurar que apenas *tweets* em inglês fossem incluídos na análise, foi implementada uma filtragem de idioma. Esse processo utilizou a biblioteca *detect* para detecção de caracteres para identificar padrões que indicassem idiomas distintos do inglês. *Tweets* contendo caracteres fora do conjunto UTF-8, ou identificados como pertencentes a outros idiomas, foram removidos. Esse filtro foi essencial para manter a consistência linguística dos dados e assegurar que a análise não fosse afetada por variações linguísticas.

Algoritmo 1 – Algoritmo para filtragem de textos em Inglês

```
1 def is_english(text):
2     try:
3         return detect(text) == 'en'
4     except:
5         return False
```

Fonte: O Autor (2024).

7.3 LIMPEZA DOS TEXTOS

Para enriquecer a análise dos *tweets*, foram realizadas extrações de informações específicas do texto presente na coluna *Tweet*. A extração de *hashtags* e *emojis*, elementos frequentemente utilizados em *tweets*, permite uma compreensão mais profunda dos temas e sentimentos expressos nas postagens.

Extração de Hashtags: Todas as *hashtags* foram extraídas para cada *tweet* usando expressões regulares para identificar o padrão "#palavra". O resultado foi armazenado na coluna `Hashtags` como uma lista, facilitando a análise de *hashtags* frequentes e a associação de temas entre diferentes *tweets*.

Algoritmo 2 – Algoritmo para Extração de Hashtags

```
1 def extract_hashtags(text):
2     return re.findall(r'#\w+', text)
```

Fonte: O Autor (2024).

Extração de Emojis: Utilizando uma biblioteca especializada, como a '*emoji*', foram identificados e extraídos os *emojis* presentes nos *tweets*. Esses símbolos foram armazenados na coluna `Emoji`, também como uma lista, permitindo a análise de padrões emocionais e o uso de ícones visuais como complemento do texto.

Algoritmo 3 – Algoritmo para Extração de Emojis

```
1 def extract_emojis(text):
2     return ''.join(c for c in text if c in emoji.EMOJI_DATA)
```

Fonte: O Autor (2024).

Além disso, o texto dos *tweets* foi limpo para remover menções, *hashtags*, pontuações, *links* e *emojis* preservando apenas o conteúdo textual relevante. Essa limpeza auxiliou na consistência da coluna `Tweet`, permitindo que apenas informações significativas fossem mantidas para análise posterior.

Algoritmo 4 – Algoritmo para limpeza do texto

```
1  # Fun  o para limpeza completa do texto
2  def clean_text(text):
3      # Remocao de Links
4      text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
5      # Remocao de mencoes
6      text = re.sub(r'@\w+|#\w+', '', text)
7      # Remocao de emojis
8      text = emoji.replace_emoji(text, replace='')
9      # Remo  o de quebra de linha
10     text = text.replace('\n', '')
11     # Remo  o de pontua  o do texto
12     return text.translate(str.maketrans('', '', string.punctuation))
```

Fonte: O Autor (2024).

7.4 UNIÃO DOS *DATASET*S E AJUSTES FINAIS

Após a padronização de cada *dataset* individual, a etapa final consistiu na unificação dos arquivos em um único *dataset* consolidado. Este processo foi conduzido utilizando um algoritmo personalizado para a leitura, deduplicação e organização dos dados.

7.4.1 Processo de Unificação

O procedimento de unificação dos dados coletados seguiu os seguintes passos:

1. **Identificação de Duplicatas de IDs:** Para *datasets* que possuíam a coluna `ID`, foi realizada uma verificação de duplicatas. Caso identificados IDs duplicados dentro do mesmo *dataset*, um deslocamento sequencial foi aplicado para criar novos IDs únicos, garantindo a integridade das identificações em todo o conjunto de dados.

2. **Deduplicação Baseada em Conteúdo:** Após a leitura de cada parte, foi realizada a remoção de duplicatas com base nas colunas `Date` e `Tweet`. Este método assegurou que *tweets* com textos idênticos publicados na mesma data fossem removidos, evitando redundâncias e mantendo a precisão dos dados.

3. **Ordenação e Unificação:** Os dados foram então combinados em um único *DataFrame*, aplicando ordenação cronológica com base na coluna `Date` para facilitar análises temporais subsequentes. Por fim, duplicatas residuais foram eliminadas para garantir a exclusividade de cada entrada.

Algoritmo 5 – Algoritmo para unificação de Dados e tratamento de exceções

```
1  def merge_files(file_paths, output_file):
2      ...
3      for file_path in file_paths:
4          chunk = pd.read_csv(file_path, encoding='utf-8', low_memory=False)
5          if 'ID' in chunk.columns:
6              duplicate_ids = chunk[chunk['ID'].duplicated(keep=False)]['ID']
7              if not duplicate_ids.empty:
8                  id_offset = max(seen_ids) + 1 if seen_ids else 1
9                  for idx in chunk[chunk['ID'].isin(duplicate_ids)].index:
10                     chunk.loc[idx, 'ID'] += id_offset
11                     id_offset += 1
12             seen_ids.update(chunk['ID'])
13         all_parts.append(chunk.copy())
14
15     combined_df = pd.concat(all_parts, ignore_index=True)
16     combined_df = combined_df.drop_duplicates(subset=['Date'], keep='first'
           )
17     combined_df = combined_df.drop_duplicates(subset=['Tweet'], keep='first
           ')
18     combined_df = combined_df.sort_values(by='Date', ascending=True)
19     combined_df.to_csv(output_file, index=False, encoding='utf-8')
```

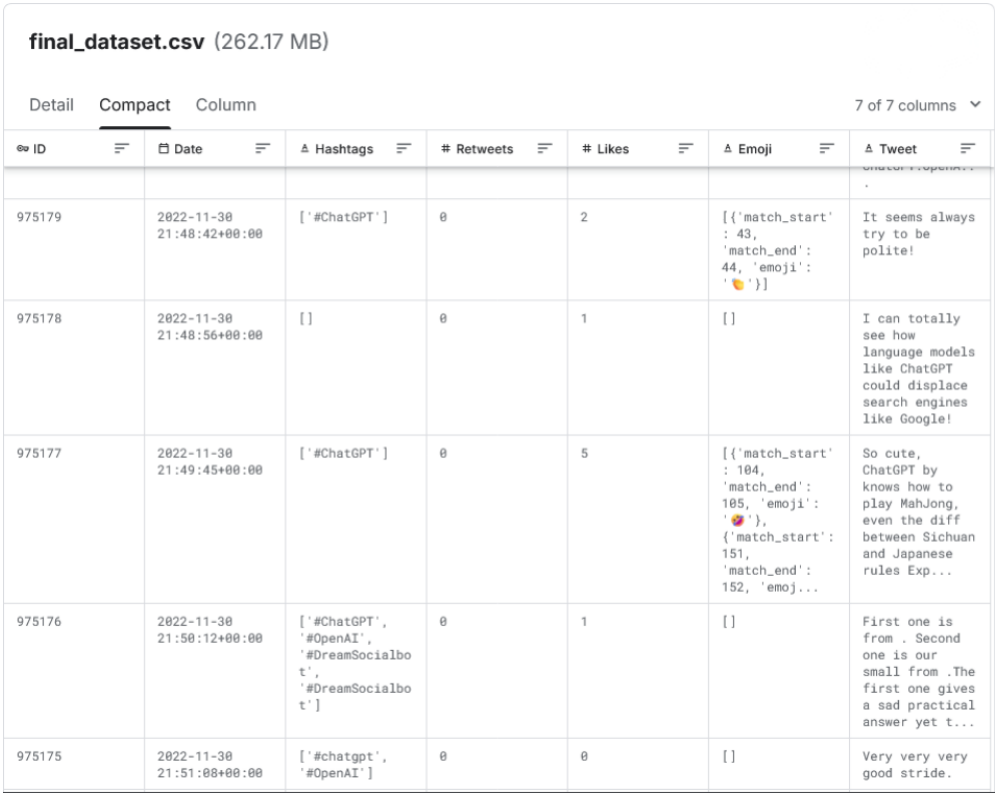
Fonte: O Autor (2024).

7.5 DISPONIBILIZAÇÃO DO *DATASET* FINAL

Após a conclusão da unificação e deduplicação, o *dataset* final foi exportado para um arquivo CSV, contendo 262MB com 1.199.923 linhas, preservando a estrutura padronizada definida anteriormente, como pode ser visto na Figura 10. Este *dataset* consolidado está agora

pronto para análises avançadas, contendo informações limpas, organizadas e livres de duplicatas. Esse formato final garante integridade e minimiza ruídos textuais e redundâncias, permitindo maior eficácia em análises mais simples.

Figura 10 – Dataset Final



final_dataset.csv (262.17 MB)

Detail | Compact | Column — 7 of 7 columns

ID	Date	Hashtags	# Retweets	# Likes	Emoji	Tweet
						.
975179	2022-11-30 21:48:42+00:00	['#ChatGPT']	0	2	[{'match_start': 43, 'match_end': 44, 'emoji': '🍋'}]	It seems always try to be polite!
975178	2022-11-30 21:48:56+00:00	[]	0	1	[]	I can totally see how language models like ChatGPT could displace search engines like Google!
975177	2022-11-30 21:49:45+00:00	['#ChatGPT']	0	5	[{'match_start': 104, 'match_end': 105, 'emoji': '🀄'}, {'match_start': 151, 'match_end': 152, 'emoj...	So cute, ChatGPT by knows how to play MahJong, even the diff between Sichuan and Japanese rules Exp...
975176	2022-11-30 21:50:12+00:00	['#ChatGPT', '#OpenAI', '#DreamSocialbot', '#DreamSocialbot']	0	1	[]	First one is from . Second one is our small from .The first one gives a sad practical answer yet t...
975175	2022-11-30 21:51:08+00:00	['#chatgpt', '#OpenAI']	0	0	[]	Very very very good stride.

Fonte: O Autor(2024).

Ao final o conjunto de dados foi disponibilizado na plataforma *Kaggle*, intitulado ***IA Tweets X***[1], com o objetivo de oferecer um recurso estruturado e acessível para a comunidade acadêmica e interessados na área de Inteligência Artificial e redes sociais. O *dataset* foi publicado com as seguintes *tags*: *Social Networks*, *Artificial Intelligence*, *English*, *Text Classification* e *Classification*, destacando sua relevância para análises textuais, classificações e estudos relacionados à IA.

A descrição da página no *Kaggle* inclui uma explicação detalhada da origem dos dados, que foram consolidados a partir de três *datasets* disponíveis na mesma plataforma: o *ChatGPT Launch Discussions on X*, desenvolvido por *Waqas Aman* e *Aqdas Malik*; o *100K Tweets about ChatGPT*, criado por *Anil*; e o *Chatgpttweets*, produzido por *Nora Abdo*, estes dados estão bem claros na descrição para promover a transparência e o respeito às contribuições prévias. A estrutura do *dataset* foi cuidadosamente normalizada para incluir campos como ID, data, texto do *tweet*, *hashtags*, *retweets*, *likes* e *emojis*, garantindo uma organização clara e adequada para uso em análises posteriores.

[1] Disponível em:https://www.kaggle.com/datasets/lgnachtigall/ia-tweets-from-x

7.6 VALIDAÇÃO DO *DATASET*

Para assegurar que o *dataset* estava adequado para publicação e análise, foram realizadas simulações que avaliaram sua usabilidade em dois contextos principais: o pós-processamento textual e a aplicação de análises de sentimentos utilizando diferentes abordagens.

7.6.1 Pós-Processamento dos Dados

O pós-processamento textual foi realizado com o objetivo de otimizar os dados para análises posteriores, reduzindo ruídos e padronizando as entradas. Esta etapa incluiu a remoção de *stopwords* — palavras sem valor semântico significativo, como artigos e preposições —, e a lematização, que consiste em reduzir palavras à sua forma base, mantendo o significado original. Essas operações foram implementadas utilizando as ferramentas disponibilizadas pelas bibliotecas `nltk`, em particular os módulos *stopwords* e *WordNetLemmatizer*. O resultado foi um conjunto de textos mais conciso e coerente, facilitando a aplicação de técnicas analíticas avançadas.

Algoritmo 6 – Algoritmo para remover stopwords

```
def remove_stopwords(text):
    new_text = []
    for word in text.split():
        if word.lower() not in stopword:
            new_text.append(word)
    return "␣".join(new_text)
```

Fonte: O Autor (2024).

Algoritmo 7 – Algoritmo para lematização

```
def preprocess_text(text):
    text = remove_stopwords(text)# Remover stopwords
    tokens = word_tokenize(text) # Tokenizar as palavras
    # Lematizar as palavras
    lemmatized_tokens = [wordnet_lemmatizer.lemmatize(word, pos='v') for
        word in tokens]
    # Retornar as palavras lematizadas de volta em uma string
    return "␣".join(lemmatized_tokens)
```

Fonte: O Autor (2024).

7.6.2 Análise de Sentimentos com VADER

A análise de sentimentos foi conduzida inicialmente utilizando o modelo VADER, disponível na biblioteca `nltk`, sendo uma ferramenta baseada em uma estrutura léxica especializada em identificar polaridades emocionais em textos curtos, como *tweets*. Este modelo é

amplamente utilizado para tarefas de análise de sentimentos devido à sua simplicidade e rapidez, sendo especialmente adequado para textos curtos, como *tweets*. No Algoritmo 8 mostra como por meio do módulo `SentimentIntensityAnalyzer`, passando o texto como entrada ele retorna uma pontuação de -1 a 1, e com isso cada entrada do *dataset* foi classificada em uma das três categorias: positiva, negativa ou neutra, sendo positiva a pontuação superior a 0.05, negativa inferior a -0.05 e entre estes valores neutra como pode ser visto no Algoritmo 9. A aplicação dessa abordagem confirmou que o *dataset* estava bem estruturado para análises rápidas de sentimento.

Algoritmo 8 – Algoritmo para análise de sentimentos Vader

```
1 def analyze_sentiment(text):
2     analyzer = SentimentIntensityAnalyzer()
3     sentiment_score = analyzer.polarity_scores(text)
4     return sentiment_score
```

Fonte: O Autor (2024).

Algoritmo 9 – Algoritmo para classificar os sentimentos Vader

```
1 def classificar_sentimento(compound_score):
2     if compound_score >= 0.05:
3         return 'Positivo'
4     elif compound_score <= -0.05:
5         return 'Negativo'
6     else:
7         return 'Neutro'
```

Fonte: O Autor (2024).

7.6.3 Análise de Sentimentos com BERT

Complementarmente, uma análise de sentimentos mais avançada foi realizada utilizando o modelo baseado na arquitetura BERT, explorando uma técnica de *transfer learning*. Essa abordagem aproveita conhecimentos previamente adquiridos para novas tarefas de NLP, facilitando a aplicação em contextos distintos. O modelo selecionado foi o *nlptown/bert-base-multilingual-uncased-sentiment*, disponível na biblioteca *Hugging Face Transformers*, especializado em análise de sentimentos multilíngue. A implementação utilizou *Python*, integrando as bibliotecas *transformers* e *torch*.

Para utilizar o modelo, é necessário configurar o *tokenizer* e carregar o modelo pré-treinado conforme indicado no Algoritmo 10:

Algoritmo 10 – Algoritmo para análise de sentimentos BERT

```
1 # Inicializar o tokenizer e o modelo
2 model_name = 'nlptown/bert-base-multilingual-uncased-sentiment'
3 tokenizer = BertTokenizer.from_pretrained(model_name)
```

```
model = BertForSequenceClassification.from_pretrained(model_name)

# Configurar o modelo para o modo de avalia   o
model.eval()

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model = model.to(device)
```

Fonte: O Autor (2024).

Após a configuração inicial, o modelo é utilizado para processar entradas textuais e produzir uma classificação de sentimento. O texto é tokenizado, truncado para atender ao limite máximo de *tokens* do modelo (512 *tokens*), e processado para obter as probabilidades associadas às classes de sentimento. O método `get_bert_embeddings`, conforme demonstrado no algoritmo 11, realiza a tokenização do texto de entrada, a inferência e a extração do rótulo de sentimento correspondente

Algoritmo 11 – Algoritmo para Tokenizar e classificar os sentimentos BERT

```
def get_bert_embeddings(text):
    inputs = tokenizer(text, return_tensors='pt', truncation=True, padding=
        True, max_length=512).to(device)
    with torch.no_grad():
        outputs = model(**inputs)

    logits = outputs.logits.squeeze().cpu()
    probabilities = torch.nn.functional.softmax(logits, dim=-1)
    predicted_class = torch.argmax(probabilities).item()
    confidence_score = probabilities[predicted_class].item()
    sentiment_labels = ['Negativo', 'Negativo', 'Neutro', 'Positivo', '
        Positivo']
    sentiment_label = sentiment_labels[predicted_class]

    return sentiment_label, confidence_score
```

Fonte: O Autor (2024).

O resultado inclui o rótulo de sentimento (negativo, neutro ou positivo) e o nível de confiança da classificação. Essa abordagem foi aplicada ao *dataset*, categorizando os textos em três classes principais — positiva, negativa e neutra —, compatíveis com as categorias utilizadas anteriormente.

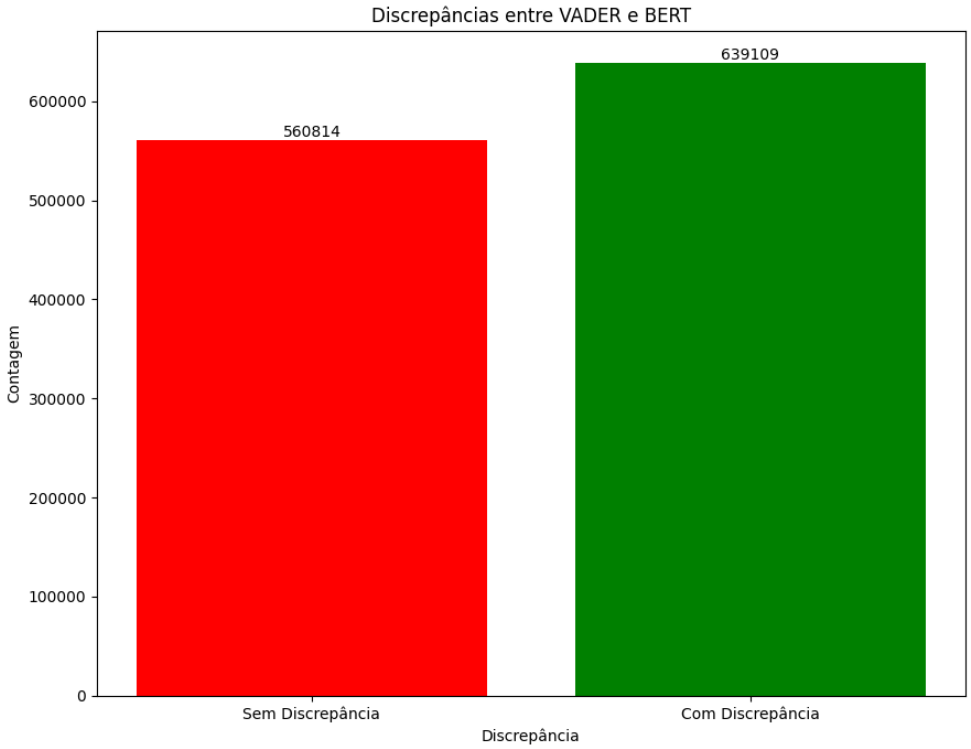
O uso do modelo BERT exigiu maior capacidade computacional devido ao alto número de parâmetros da arquitetura. No entanto, os resultados obtidos foram consistentes com as análises realizadas por métodos alternativos, evidenciando a qualidade do *dataset* e a aplicabilidade da metodologia para estudos relacionados a sentimentos e opiniões. A análise também destacou a viabilidade de modelos *pre-trained* para exploração de grandes volumes de dados textuais,

como os oriundos de redes sociais.

7.6.4 Comparação entre os Modelos

A comparação entre os modelos VADER e BERT revelou diferenças significativas tanto no desempenho computacional quanto nos resultados das classificações. O modelo VADER, conhecido por sua eficiência, processou o *dataset* em aproximadamente 120 segundos, enquanto o BERT, devido à sua arquitetura mais complexa, demandou cerca de 11 horas para concluir a análise. Essa discrepância no tempo de execução reflete o trade-off clássico entre rapidez e sofisticação no campo do Processamento de Linguagem Natural.

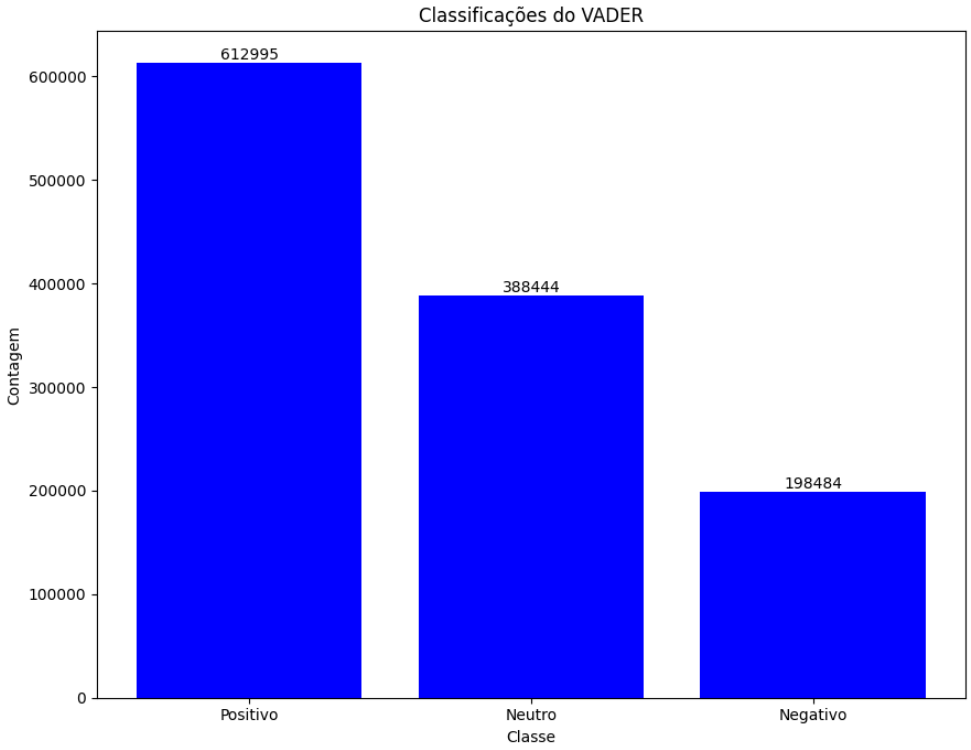
Figura 11 – Distribuição das discrepâncias entre os modelos VADER e BERT.



Fonte: O Autor(2024).

A Figura 11 ilustra a discrepância entre os dois modelos, destacando que, apesar das diferenças metodológicas, a maior parte das classificações foi consistente entre ambos (53,3%). Ainda assim, 46,7% das amostras apresentaram discordâncias, o que indica nuances que o modelo VADER pode ter perdido devido à sua abordagem baseada em léxicos.

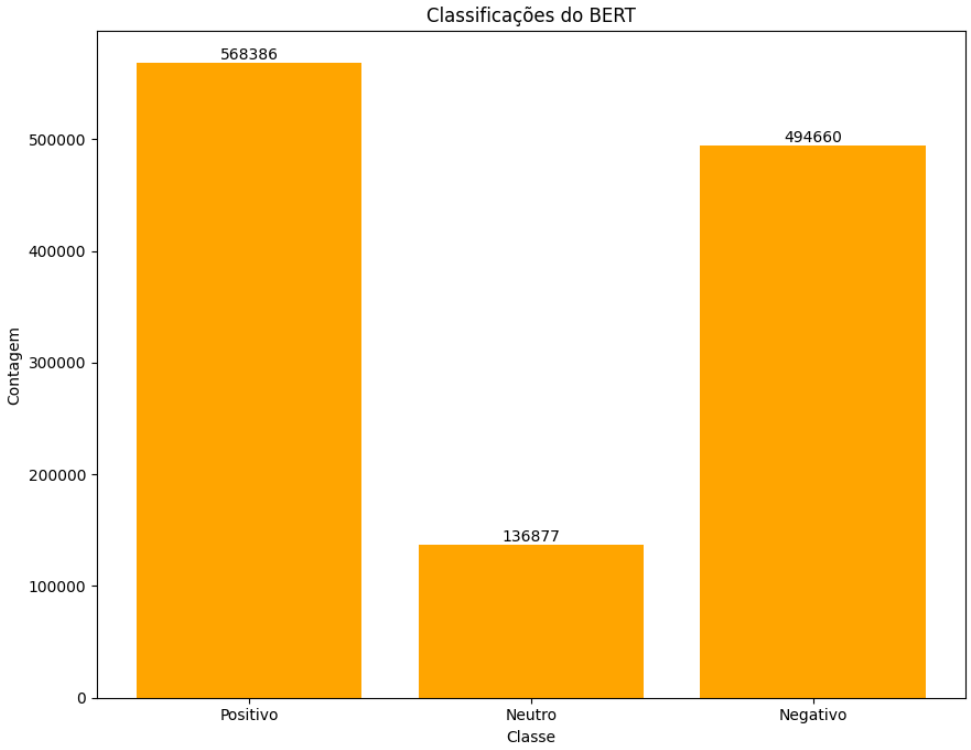
Figura 12 – Distribuição das classificações utilizando o modelo VADER.



Fonte: O Autor(2024).

A Figura 12 mostra a distribuição das classificações do modelo VADER: 51,1% das amostras foram identificadas como positivas, 32,4% como neutras e 16,5% como negativas. Esses resultados são compatíveis com a natureza do *dataset*, sugerindo uma maior proporção de textos com tonalidade positiva.

Figura 13 – Distribuição das classificações utilizando o modelo BERT.



Fonte: O Autor(2024).

Já a Figura 13 evidencia as classificações realizadas pelo modelo BERT: 47,4% positivas, 11,4% neutras e 41,2% negativas. A maior proporção de textos negativos em comparação com o VADER sugere que o BERT conseguiu capturar nuances semânticas mais profundas que indicam emoções negativas, frequentemente ignoradas por abordagens léxicas simplificadas.

Além disso, o BERT apresentou maior capacidade de identificação de polaridades extremas, possivelmente devido à sua habilidade de interpretar contexto de forma mais robusta. Em contraste, a abordagem léxica do VADER mostrou limitações na detecção de sentimentos mais complexos, como ironia e sarcasmo.

Em resumo, enquanto o VADER se mostrou eficiente para análises rápidas e menos complexas, o BERT proporcionou resultados mais detalhados e, potencialmente, mais precisos, ao custo de um maior investimento computacional. A escolha entre os modelos dependerá do objetivo específico e das restrições de recursos.

8 CONSIDERAÇÕES FINAIS

Primariamente o projeto enfrentou limitações impostas pela mudança da API do *Twitter* para o *X*, que inviabilizou a extração direta de dados da rede social durante o desenvolvimento. Essa alteração restringiu o acesso a informações diretamente da plataforma, tornando necessário o uso de *datasets* previamente coletados e compartilhados por terceiros. Essa dificuldade destacou a importância de trabalhar com fontes confiáveis e éticas para obter os dados necessários, além de reforçar a relevância de adaptações metodológicas frente a alterações inesperadas em ferramentas tecnológicas.

Um dos principais desafios enfrentados foi a diversidade de idiomas e caracteres presentes nos textos dos *tweets*. Como os dados foram coletados de diferentes regiões e continham informações de usuários de várias partes do mundo, o conteúdo incluía caracteres e símbolos específicos de diferentes idiomas. Outro aspecto que exigiu atenção foi a limpeza dos dados textuais, para facilitar análises mais claras, junto a isto a remoção de links, menções e pontuações dos *tweets*, além da extração de *hashtags* e *emojis*. Essas etapas ajudaram a simplificar o texto e a isolar o conteúdo relevante, removendo elementos que poderiam comprometer a análise sem acrescentar informações significativas. O resultado foi um *dataset* mais limpo e estruturado, composto por *tweets* padronizados e representativos do conteúdo, e adicionalmente, para evitar inconsistências, os *datasets* foram tratados em relação à duplicidade de *tweets*, onde registros sobrepostos foram identificados e removidos para assegurar a exclusividade dos dados no *dataset* final.

A validação do *dataset* demonstrou sua adequação tanto para análises básicas de sentimentos, como aquelas realizadas com o modelo VADER, quanto para abordagens mais avançadas, como as conduzidas com o modelo BERT. As simulações de pós-processamento textual, que incluíram técnicas de remoção de *stopwords* e lematização, foram fundamentais para otimizar os dados para as análises subsequentes, além de demostrarem algumas das manipulações aos quais os dados, ainda, podem sofrer de acordo com a análise final desejada. Adicionalmente, a utilização de diferentes ferramentas, como `nltk` e `transformers`, mostrou que o *dataset* está preparado para atender a uma ampla gama de necessidades analíticas.

Ao final, o conjunto de dados foi disponibilizado na plataforma *Kaggle*, intitulado ***IA Tweets X***[1], com o objetivo de oferecer um recurso estruturado e acessível para a comunidade acadêmica e interessados na área de Inteligência Artificial e redes sociais. O *dataset* foi publicado com as seguintes *tags*: *Social Networks*, *Artificial Intelligence*, *English*, *Text Classification* e *Classification*, destacando sua relevância para análises textuais, classificações e estudos relacionados à IA.

[1] Disponível em: https://www.kaggle.com/datasets/lgnachtigall/ia-tweets-from-x

Como resultado desse trabalho, o *dataset* final foi disponibilizado publicamente na plataforma *Kaggle*, sob o título ***IA Tweets X***[2]. Essa disponibilização visa facilitar o acesso ao conjunto de dados por parte da comunidade acadêmica e de pesquisadores interessados na relação entre inteligência artificial e redes sociais. A publicação foi acompanhada por *tags* relevantes, como *Social Networks*, *Artificial Intelligence*, *English*, *Text Classification* e *Classification*, destacando seu potencial para análises textuais, estudos de classificação e investigações mais amplas no campo da IA.

De maneira geral, este trabalho abordou com sucesso os desafios inerentes à unificação e padronização de *datasets*, culminando na criação de um recurso robusto e versátil. Ele não apenas demonstrou a importância de um rigoroso tratamento de dados para garantir a qualidade e a consistência das informações, mas também forneceu uma base significativa para investigações futuras que explorem as interações entre inteligência artificial e o universo das redes sociais.

8.1 TRABALHOS FUTUROS

Como continuidade para trabalhos futuros, alternativas como a expansão da análise de sentimentos para incluir modelos adicionais, tanto de natureza mais simples quanto para a classificação de sentimentos mais complexos. Modelos baseados em arquiteturas como GPT, XLNet ou RoBERTa poderiam ser explorados para avaliar aspectos mais sutis, como sarcasmo e ironia, ou mesmo emoções específicas, como raiva e alegria.

Além disso, há a possibilidade de testar diferentes técnicas de pré-processamento textual, como a stemização ou a aplicação de outros métodos de normalização, para avaliar seu impacto na qualidade das análises. O uso de combinações de abordagens, que incluam tanto métodos lexicais quanto baseados em aprendizado profundo, também pode trazer novas perspectivas.

[2] Disponível em: https://www.kaggle.com/datasets/lgnachtigall/ia-tweets-from-x

REFERÊNCIAS

ABREU, F. S. G. da Gama e. Desmistificando o conceito de etl. **Revista de Sistemas de Informação**, n. 02, 2008. Disponível em: <https://www.fsma.edu.br/si/Artigos/V2_Artigo1.pdf>. Acesso em: Jun 2024.

ALJEDANI, W. *et al.* Sentiment analysis optimization using vader lexicon on machine learning approach. **IEEE Xplore**, 2022. Disponível em: <https://ieeexplore.ieee.org/document/9855341>. Acesso em: 17 nov. 2024.

ANACONDA. 2022 state of data science. 2022. Disponível em: <https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022>.

BANIK, R. **The Complete Pokemon Dataset**. 2017. Disponível em: <https://www.kaggle.com/datasets/rounakbanik/pokemon>. Acesso em: Abr 2024.

BROUCKE, S. vanden; BAESENS, B. **Practical Web Scraping for Data Science**: Best practices and examples with python. [S.l.]: Apress Berkeley, CA, 2018.

CANALTECH. **Facebook**. [S.l.], 2024. Disponível em: <https://canaltech.com.br/empresa/facebook/>.

____. **Twitter**. [S.l.], 2024. Disponível em: <https://canaltech.com.br/empresa/twitter/>.

CARDONE, H. T. M.; MARINH, I. J. P.; VAZ, G. J. Uma metodologia para criação de stop lists em sistemas de recuperação de informação em domínios específicos. 2012. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/80003/1/19.pdf>. Acesso em: Jun 2024.

CRISTESCU, M. *et al.* Applying bert and vader in hr sentiment analysis. **Creative Space Association Journal**, p. 6–23, 2023.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805v2 [cs.CL]**, 2019. Disponível em: <https://arxiv.org/abs/1810.04805v2>.

DOMO. Data transformation techniques, types, and methods. 2024. Disponível em: <https://www.domo.com/learn/article/data-transformation-techniques>.

DURGA, S. V. *et al.* Mining social media data for sentiment analysis and trend prediction. v. 10, p. 1557–1562, 2023.

FERNANDES, H. M. O que é uma api e para que serve? 2020. Disponível em: <https://marquesfernandes.com/tecnologia/o-que-e-uma-api-e-para-que-serve/>. Acesso em: Jun 2024.

FERREIRA, H. H. Processamento de linguagem natural e classificação de textos em sistemas modulares. 2019. Disponível em: <https://bdm.unb.br/bitstream/10483/25114/1/2019_HugoHondaFerreira_tcc.pdf>. Acesso em: Jun 2024.

HAMEED, M.; NAUMANN, F. Data preparation: A survey of commercial tools. **SIGMOD Record**, v. 49, n. 3, 2020.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 8, n. 1, p. 216–225, 2014. Disponível em: <https://doi.org/10.1609/icwsm.v8i1.14550>.

IBM. **ETL (extract, transform, load)**. 2020. Disponível em: <https://www.ibm.com/cloud/learn/etl>. Acesso em: Mai 2024.

JACOBSON, D.; BRAIL, G.; WOODS, D. **APIs: A Strategy Guide**. [S.l.]: O'Reilly Media, Inc., 2011. ISBN 9781449308926.

KAGGLE. **How to Use Kaggle**: Datasets. [S.l.], 2024. Disponível em: <https://www.kaggle.com/docs/datasets>. Acesso em: Abr 2024.

KAPUR, K.; HARIKRISHNAN, R. Comparative study of sentiment analysis for multi-sourced social media platforms. **arXiv preprint arXiv:2212.04688 [cs.CL]**, 2022. Disponível em: <https://arxiv.org/abs/2212.04688>.

KEMP, S. **DIGITAL 2024: BRAZIL**. [S.l.], 2024. Disponível em: <https://datareportal.com/reports/digital-2024-global-overview-report>. Acesso em: Jun 2024.

____. **DIGITAL 2024: GLOBAL OVERVIEW REPORT**. [S.l.], 2024. Disponível em: <https://datareportal.com/reports/digital-2024-brazil>. Acesso em: Jun 2024.

KHDER, M. A. Web scraping or web crawling: State of art, techniques, approaches and application. **Int. J. Advance Soft Compu. Appl**, v. 13, n. 3, 2021. Disponível em: <https://www.researchgate.net/publication/357401723_Web_Scraping_or_Web_Crawling_State_of_Art_Techniques_Approaches_and_Application>. Acesso em: Mai 2024.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. [S.l.]: John Wiley Sons, 2011. Acesso em: Jun 2024.

KIMBALL, R.; ROSS, M. **The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence**. 3. ed. [S.l.]: John Wiley Sons, 2013.

KINSTA. **O que é Web Scraping? Como Extrair Legalmente o Conteúdo da Web**. [S.l.], 2023. Disponível em: <https://kinsta.com/pt/base-de-conhecimento/o-que-e-web-scraping/>. Acesso em: Jun 2024.

LUNELLI, E. J. ExtraÇÃo e tratamento de dados do twitter no perÍodo eleitoral de 2022 para criaÇÃo de um dataset. 2022.

MATHIEN, H. **European Soccer Database**. 2016. Disponível em: <https://www.kaggle.com/datasets/hugomathien/soccer>. Acesso em: Abr 2024.

META. **facebook app**. [S.l.], 2024. Disponível em: <https://about.meta.com/technologies/facebook-app/>.

MIRANDA, G. V. d. F. *et al.* Extração e avaliação de uma base de dados sobre criminalidade em português a partir do twitter. ANAIS DO SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO UBÍQUA E PERVASIVA (SBCUP), 2023. Disponível em: <https://sol.sbc.org.br/index.php/sbcup/article/view/24995/24816>.

MITCHELL, T. M. **Machine Learning**. 2017. Disponível em: <https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>. Acesso em: mai 2024.

MOONEY, P. **Chest X-Ray Images (Pneumonia)**. 2018. Disponível em: <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>. Acesso em: Abr 2024.

MUNIZ, R. I. V. C. S. Publicação de dados abertos conectados sobre os transplantes realizados no imip. **Brazilian Symposium on Databases**, v. 33, 2018.

NIP, J. Y. M.; BERTHELIER, B. Social media sentiment analysis. **Encyclopedia**, v. 4, n. 4, p. 1590–1598, 2024. ISSN 2673-8392. Disponível em: <https://www.mdpi.com/2673-8392/4/4/104>.

OPEN KNOWLEDGE FOUNDATION. The open definition. 2024. Disponível em: <https://opendefinition.org/>. Acesso em: Abr 2024.

PEREIRA, M. **Estatísticas Twitter (X) - 2023**. 2023. Online. Disponível em: <https://definicao.marketing/estatisticas-twitter/>.

PRESS, G. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. **Forbes**, 2016. Disponível em: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>.

PROFERES, N. *et al.* Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. **Social Media + Society**, v. 7, n. 2, p. 20563051211019004, 2021. Disponível em: <https://doi.org/10.1177/20563051211019004>.

REDHAT. **APIs**. [S.l.], 2018. Disponível em: <https://www.redhat.com/pt-br/topics/api>. Acesso em: Mai 2024.

REIS, G. G. Análise de incidentes de segurança utilizando dados do twitter. 2023. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/40982/1/AnaliseIncidentesSeguran%C3%A7a.pdf>. Acesso em: Jun 2024.

REIS, J.; HOUSLEY, M. **Fundamentos de Engenharia de Dados: Projete e Construa Sistemas de Dados Robustos**. [s.n.], 2023. ISBN 8575228773, 9788575228777. Disponível em: <https://books.google.com.br/books/about/Fundamentos_de_Engenharia_de_Dados.html?id=Q1fiEAAAQBAJ&redir_esc=y>. Acesso em: Jun 2024.

SANTOS, F. A. *et al.* Processamento de linguagem natural em textosde mídias sociais: Fundamentos, ferramentas e aplicações. 2022. Disponível em: <https://books-sol.sbc.org.br/index.php/sbc/catalog/view/106/473/746>. Acesso em: Jun 2024.

SESSA, M. G. Disinformation on facebook: Research and content moderation policies. 2024.

SILVA, T. H. *et al.* Urban computing leveraging location-based social network data: A survey. **ACM Comput. Surv**, n. 17, 2019. Disponível em: <https://homepages.dcc.ufmg.br/~fabricio/download/silva-acm-surveys.pdf>. Acesso em: Jun 2024.

SOUZA, F. B. de. **Uma análise empírica de interações em redes sociais**. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2010. Disponível em: <https://repositorio.ufmg.br/handle/1843/SLSS-85BN96>.

SQLITE. **What Is SQLite?** [S.l.], 2024. Disponível em: <https://www.sqlite.org/index.html>. Acesso em: Abr 2024.

VAGHANI, N. **Chatbot dataset**. 2023. Disponível em: <https://www.kaggle.com/datasets/niraliivaghani/chatbot-dataset>. Acesso em: Abr 2024.

VIDA, E. d. S. *et al.* **Data warehouse**. Grupo A, 2021. ISBN 9786556901916. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9786556901916>. Acesso em: Mai 2024.

WENINGER, T.; ZHU, X. A.; HAN, J. An exploration of discussion threads in social news sites: A case study of the reddit community. 2013. Disponível em: <https://dejanmarketing.com/wp-content/uploads/2013/10/reddit-paper.pdf>. Acesso em: Jun 2024.

X. **Docs**. [S.l.], 2024. Disponível em: <https://developer.x.com/en/docs>.

____. **Glossary**. [S.l.], 2024. Disponível em: <https://help.x.com/en/resources/glossary>.

YAGUI, M. M. M.; MAIA, L. F. M. P. Data mining of social manifestations in twitter: An etl approach focused on sentiment analysis. ANAIS DO SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 2017. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/6019/5917>.

YOUVAN, D. C. **Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application**. 2024. ResearchGate. Disponível em: <https://www.researchgate.net/publication/381650914>.

ZHANG, X.; QI, X. Performance evaluation of reddit comments using machine learning and natural language processing methods in sentiment analysis. **arXiv preprint arXiv:2405.16810v2 [cs.CL]**, 2024. Disponível em: <https://arxiv.org/abs/2405.16810v2>.

ZHAO, B. **Encyclopedia of Big Data**. Springer, Cham, 2017. ISBN 978-3-319-32001-4. Disponível em: <https://www.researchgate.net/publication/317177787_Web_Scraping>. Acesso em: Jun 2024.