UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE FÍSICA

Carlos Pereira de Castro Neto

Extração de Dados e Análise de Sentimento:
com diferentes dicionários léxicos

São Carlos - SP
2022

Carlos Pereira de Castro Neto

Extração de Dados e Análise de Sentimento:
com diferentes dicionários léxicos

Trabalho de conclusão de curso
apresentado como requisito parcial para
obtenção do título de bacharelado em
Engenharia Física pela Universidade
Federal de São Carlos,

Prof. Dr. Fábio Aparecido Ferri
Orientador

São Carlos - SP
2022

Resumo

As redes sociais vêm sendo cada dia mais um expositor para as opiniões e sentimentos das pessoas. Eleições por todo mundo foram muito influenciadas por elas, valendo ressalvar o impacto nas eleições norte-americanas de 2016, quando o atual ex-presidente Donald Trump foi eleito, e nas eleições brasileiras de 2018, quando o atual presidente Jair Messias Bolsonaro se elegeu.

Com a evolução dos conceitos e ferramentas associados ao processamento e armazenamento de dados, pode-se analisar grandes volumes e variedades de dados de forma a buscar algum valor para eles.

Diversas áreas que utilizam de Inteligência Artificial e outras estratégias vem surgindo, com o objetivo de unir a oportunidade de grande material para análise e uma extensa gama de ferramentas para tal, como o processamento de linguagem natural. Um ramo especifico dessa área é a análise de sentimentos, que nos provê parte ou alguma consideração sobre textos.

Almejando conhecer esse processo e disponibilizar material, foi feita uma análise de sentimentos aos tweets durante e após da entrevista do presidenciável Luiz Inácio Lula da Silva no canal CNN Brasil, no dia 13 de setembro de 2022.

As análises não geraram considerações de grande relevância para o fato, mas para o entendimento de como a área de Análise de Sentimentos é complexa e tem espaço para melhorias.

Palavras-chaves: Big Data. Python. Análise de Sentimentos. Twitter. Dicionário Léxico.

Lista de Figuras

Sumário

1. Introdução

Segundo a pesquisa TIC Domicílios 2021, divulgada em julho de 2022 pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), o percentual de residências com acesso à internet subiu de 71% para 82% no período de dois anos (2019 - 2021) [1].

Com o passar natural do tempo e devido ao isolamento social ocasionado pela pandemia do Covid-19, nos últimos anos cada vez mais pessoas vem utilizando de redes sociais para expressarem suas opiniões, sentimentos e expectativas. A partir de pesquisas na internet, é possível encontrar uma quantidade muito grande de informações sobre algum tema especifico, porém essa informação na maioria das vezes se encontra desestruturada e dispersa, sendo difícil fazer uma análise a respeito.

A área de Análise de Sentimentos ou Mineração de Opinião visa descobrir, quantificar e qualificar computacionalmente opiniões e seus conceitos relacionados, com objetivo de avaliar o sentimento sobre um determinado produto, analisar empresas na bolsa de valores, analisar sentimento sobre determinadas pessoas, entre outros [2].

O Twitter é uma rede social que permite aos usuários publicar e acompanhar atualizações pessoais dos contatos. Tais atualizações, textos com limite de 280 caracteres (desde 2018), são conhecidos como *tweets*. Desde o seu princípio a rede é utilizada pelos usuários para expressar opiniões e sentimentos. De acordo com estudos, pesquisadores têm investido cada vez mais na Análise de sentimentos no Twitter [3].

Na última década o termo *Big Data* vem ganhando força no ramo de Tecnologia da Informação (TI). *Big Data* se refere a grandes conjuntos de dados, os quais precisam devem ser processados e/ou armazenados. Dessa forma, o Twitter pode ser utilizada como uma fonte potente de *Big Data*.

O Twitter possui uma API (*Application Programming Interface*) que facilita a extração de *tweets* em tempo real e historicamente. Na versão mais recente Twitter API v2, existem 3 versões de licença para extração de dados. Para este trabalho foi utilizada a versão *Essential*, que permite a extração de 500 mil tweets por mês, numa taxa de 300 requisições a cada 15 minutos e 100 tweets por requisições, totalizando 30000 tweets a cada 15 minutos. Esse número permite a formação de uma base de dados consistente para análises.

As eleições 2022 acontecem no dia 02 de outubro de 2022 e, caso houver, segundo turno no dia 30 de outubro. Segundo o DataSenado [4], as redes sociais impactaram voto de 45% da população, de forma a serem usadas como fonte de informação para decisão do voto. Durante todo o período pré-eleições cada evento tem impacto na opinião e popular e causa uma reação na população eleitoral.

Neste trabalho iremos analisar o sentimento em *tweets* feitos durante e após a entrevista do presidenciável Luiz Inácio Lula da Silva na CNN Brasil (*Cable News Network)* no dia 13 de setembro de 2022 às 20 horas, horário de Brasília, a fim de avaliar o desempenho de diferentes dicionários léxicos em PT-BR a nível de palavras.

1.1 Objetivos

Este trabalho tem como objetivo apresentar a extração de dados da rede social Twitter e analisar o sentimento relacionado a eles para diferentes dicionários léxicos disponíveis na internet, utilizando de ferramentas da linguagem Python.

Além do Capítulo 1, introdutório, este trabalho contém outros quatro capítulos. No Capítulo 2, será apresentada a fundamentação teórica para compreensão do trabalho.

No Capítulo 3, os materiais e métodos utilizados para a extração dos dados e para as análises.

Por fim, no Capítulo 4, serão feitas as conclusões e sugestões para trabalhos futuros.

1.2 Outros Trabalhos

No ano de 2019, Gabriel Silva Monteles usou do *framework* Apache Ignite para comparar resultados produzidos por diferentes dicionários léxicos, numa análise de sentimentos a nível de sentença. Tal análise foi produzida em uma base de 2564 *tweets* relacionados a uma série de focos de incêndio que ocorreram na região amazônica do Brasil no mês de setembro daquele ano [5]. Em comparação, neste trabalho serão analisados cerca de 50 mil *tweets*.

No ano de 2017, Ana Carolina Bras Costa usou do *framework* Apache Ignite para análise de sentimento a nível de sentença, utilizando de um dicionário léxico para determinar a polaridade de *tweets* a respeito do relançamento do Super Nintendo naquele ano. Foram analisados manualmente 94 *tweets*, onde 62 foram classificados como positivos e 32 como negativos. Como resultado da aplicação, 63 *tweets* foram classificados positivos e 31 como negativos, aonde apenas 43 eram realmente positivos e 16 realmente negativos [6].

2. Fundamentos Teóricos

Neste capítulo será apresentado a fundamentação teórica necessária para o desenvolvimento e entendimento deste estudo. Redes sociais e seus conceitos, *Big Data*, Python e suas principais bibliotecas para análises e as definições de Análise de Sentimentos.

2.1 Redes Sociais

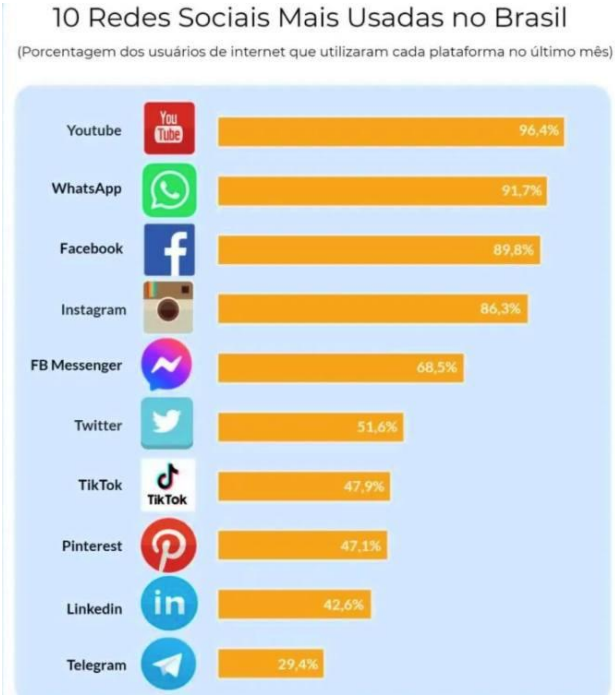
Uma rede social pode ser definida como um espaço onde ocorre um conjunto de relações entre indivíduos, grupos ou organizações, as quais partilham interesses, geralmente através de

plataformas disponíveis na Internet.

Um *Social Networking Site* (SNS) tem como objetivo disponibilizar para as pessoas um espaço próprio, o qual podem colocar informações pessoais e se relacionarem com outros usuários da mesma rede. O propósito principal não é conhecer pessoas estranhas, mas sim permitir a conexão com pessoas que já fazem parte de sua vida social [9].

Existem dezenas de redes sociais muito conhecidas no mundo. A Figura 1, mostra as redes sociais mais usadas no Brasil em mês de agosto de 2021:

Figura 1. Número de usuário nas redes sociais mais usadas no Brasil em setembro de 2021.



Fonte: [8].

Com o passar dos anos, as redes sociais evoluíram, trazendo novas funcionalidades, novas formas de interação com os outros usuários, novas formas de compartilhar informações em tempo real (*instanting messaging*). Atualmente é possível compartilhar todo o tipo de dados, texto, fotos, vídeos e até mesmo geolocalização.

As informações disponibilizadas, que antes não pareciam ter relevância vendo sendo cada vez mais alvo de estudos e valorizada. Seja para pesquisas de opinião sobre produtos ou fatos, identificação de comunidades de interesse, ou sondagem de opinião pública outrora feitas por chamadas telefônicas ou entrevistas.

2.1.1 O Twitter

O Twitter foi uma rede social inovadora a qual teve um sucesso inicial muito grande. Surgiu em 2006 permitindo aos usuários partilharem informação, criando e compartilhando

mensagem (*tweets*) de até 140 caracteres (280 caracteres a partir de 2018) por meio do *website*, por SMS e por softwares específicos.

Esses *tweets* são exibidos no perfil do usuário em tempo real e também mostradas a outros usuários seguidores. Ao contrário de outras redes, como o Facebook, o Twitter foca-se nas mensagens compartilhadas entre os usuários, de forma que o perfil do usuário passa a se tornar algo secundário. Alguns diferenciais do Twitter são o uso das "hashtags". Essas palavras-chave precedidas do caractere #, indicam que o tweet faz referência a um certo tópico. Dessa forma os "*Top Trendings"* têm fácil acesso para todos os usuários. Os usuários podem concordar ou discordar de algum *tweet* com a utilização da função "*retweetar*", replicando a mensagem e comentando-a. Alguns fatos recentes sobre o Twitter [7]:

- 1.3 bilhão de contas;
- 211 de milhões de tweets;
- 500 milhões de tweets postados todos os dias;
- No Brasil em janeiro de 2022, 19,05 milhões de brasileiros acessaram o Twitter.

2.2 Big Data

Embora um conceito novo, desde da década de 60 e 70 a humanidade começa a armazenar grandes conjuntos de dados, com os primeiros data centers e o desenvolvimento do banco de dados relacional. A primeiras redes sociais a nível global, como Facebook e Youtube, evidenciaram a quantidade de dados gerados pelos usuários na internet.

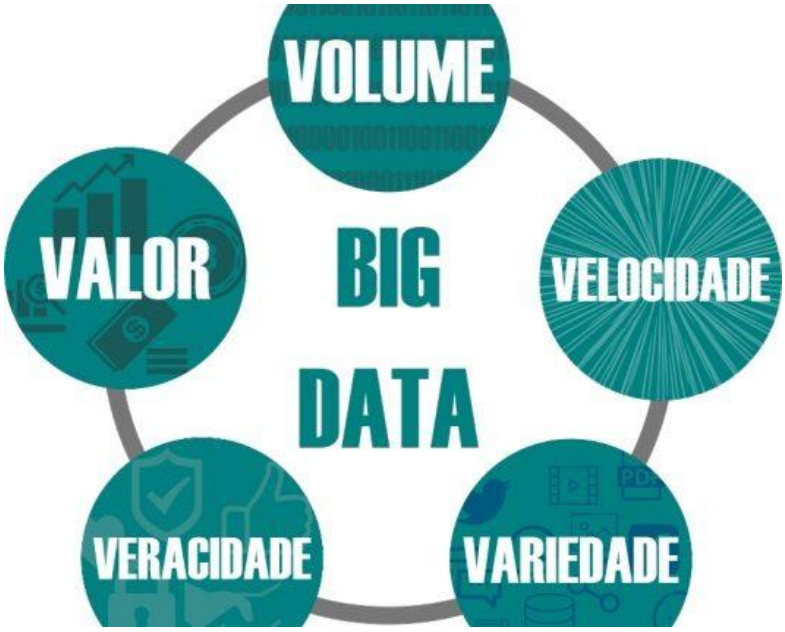
O Hadoop [11], uma estrutura de código aberto criada para armazenar e analisar grandes conjuntos, NoSQL, que se refere a tipos não relacionais de banco de dados [30], começaram a ganhar popularidade a partir do ano 2005. Mais recentemente o Apache Spark, um mecanismo de análise para processamento de dados em grade escala [17], ganhou relevância nesse cenário.

A definição de *Big Data* são dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior [12]. Estes são os três Vs. De forma mais simples, *Big Data* é um conjunto de dados maior e mais complexo, originário de novas fontes. Assim, softwares e métodos tradicionais de processamento e armazenamento não conseguem gerenciá-los, no entanto, esses conjuntos podem ter grande valor.

2.2.1 Características do Big Data

O Big Data tem como base os três Vs: volume, variedade e velocidade. Esse conceito, após reformulação, deu espaços a outros dois valores: veracidade e valor [14]. Na Figura 2 podemos observar essas características do *Big Data*.

Figura 2. Os 5 Vs do Big Data.



Fonte: [13].

- **Volume:** a característica mais importante. Refere-se ao grande volume de informações disponibilizados por redes sociais, sites, dispositivos inteligentes, empresa, entre outros.

- **Velocidade:** a taxa de geração e transmissão desses dados.

- **Variedade:** a diversidade do formato do dado. Existem três tipos de dados: dados estruturados, não estruturados e semiestruturados.

- **Veracidade:** devido ao grande volume é fácil receber dados de fontes não confiáveis ou dados que não estejam corretos. Essa característica se refere a qualidade dos dados.

- **Valor:** após a extração e processamento, empresas, instituições e pesquisadores utilizam do dado para algum fim. Nesse processo é agregado um valor mensurável para aqueles dados brutos.

2.2.2 Etapas Big Data

A cadeia de valor do *Big Data* são as etapas necessárias para gerar valor e a informação necessária para a tomada de decisão a partir desses dados. Essa cadeia, apresentada na Figura 3, possuiu quatro etapas primordiais: geração de dados, coleta de dados, armazenamento de dados e análise de dados [15].

Figura 3. Cadeia de valor Big Data.

Fonte: Elaborada pelo autor.

- **Geração de Dados:** corresponde ao processo de produção da informação. Sensores, câmeras, redes sociais, smartphones e quaisquer outros dispositivos conectados à internet.

- **Coleta de Dados:** a etapa de extrair, transformar e carregar os dados aonde são utilizadas as ferramentas e metodologias de processamento de dados.

- **Armazenamento de Dados:** nessa etapa, o acesso rápido das aplicações aos dados deve ser prezado. São utilizados bancos de dados relacionais e não-relacionais.

- **Análise de Dados:** com os dados trabalhados e armazenadas a etapa final agrega o valor ao dado, utilizando-os nas áreas de interesse pré-determinadas no projeto.

2.3 Python para Dados

Python (Figura 4) é uma linguagem *Open-Source* de propósito geral, bastante usada em data science, machine learning, desenvolvimento web, desenvolvimento de aplicativos, automação entre outros. Sendo uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional de tipagem dinâmica e forte. Foi lançada por Guido Van Rossum em 1991 [16].

Figura 4. Python logo.

Fonte: [16].

Tanto para extração quanto para análise de dados, Python é amplamente utilizada por ser uma linguagem flexível, fácil de usar e por ser de código aberto, ter uma comunidade ampla e muito ativa, e vários bibliotecas (*packages*) que complementam a linguagem e trazem novas funcionalidades e integrações.

2.3.1 PySpark

PySpark (Figura 5) é uma biblioteca que interfaceia do Apache Spark em Python. Permite que sejam escritos aplicativos Spark usando as APIs do Python e também fornece o shell PySpark para analisar interativamente os dados em um ambiente distribuído. O PySpark suporta maioria dos recursos do Spark, como o Spark SQL, Spark DataFrame, Streaming, MLlib (Machine Learning) e Spark Core [17].

Figura 5. Apache PySpark Logo.

Fonte: [17].

- **Spark SQL e DataFrame:** Spark SQL é um módulo Spark para processamento de dados estruturados. Fornece uma abstração de programação chamada DataFrame e também atua como mecanismo de consulta SQL distribuído.

- **API do Pandas no Spark:** caso o usuário tenha familiaridade com o Pandas, permite que o início imediato sem curva de aprendizado. Garante uma única base de código que funciona tanto com pandas (para testes, conjuntos de dados menores) quanto com o Spark (conjuntos de dados distribuídos).

- **Streaming:** o recurso de streaming do Apache Spark permite aplicativos interativos e analíticos poderoso em dados de streaming e históricos, enquanto herda as características de facilidade de uso e prevenção a falhas do Spark.

- **MLlib:** é uma biblioteca de aprendizado de máquina escalável que fornece um conjunto uniforme de APIs de alto nível que ajudam os usuários a criar e ajustar pipelines de aprendizado de máquina

- **Spark Core:** é o mecanismo de execução geral subjacente para a plataforma Spark sobre o qual todas as outras funcionalidades são construídas. Fornece um RDD (*Resilient Distributed DataSet*) e recursos de computação na memória.

2.3.2 Numpy e Pandas

Numpy (Figura 6) é uma biblioteca para Python que adiciona suporte para matrizes multidimensionais, juntamente com uma grande coleção de funções matemáticas de alto nível para operar nessas matrizes [18].

Figura 6. Numpy Logo.

Fonte:[18].

Pandas (Figura 7) é uma ferramenta de manipulação de dados de alto nível que é construída no pacote Numpy. A estrutura chave no Pandas é chamada de DataFrame. Os DataFrames são incrivelmente poderosos, pois permitem armazenar e manipulares dados de forma tabular em linhas e colunas [19].

Figura 7. Pandas logo.

Fonte:[19].

2.3.3 Matplotlib

O Matplotlib (Figura 8) é uma biblioteca em Python com a função de criar gráficos 2D. tem uma variedade de gráficos e diferentes formas de construção em conjuntos com diversas bibliotecas para análises de dados entre outros formatos.

Figura 8. Matplotlib Logo.

Fonte: [21].

2.3.4 Scikit Learn

Scikit-Learn (Figura 3), ou apenas SKLearn, é uma biblioteca de Machine Learning de código aberto para Python, construída baseada nas bibliotecas NumPy, SciPy e Matplotlib. Ela fornece muitos algoritmos de aprendizado supervisionado e não supervisionado, com funcionalidades como regressão, classificação, clustering, seleção de modelo [20].

Figura 9. Scikit-learn logo.

Fonte: [20].

2.3.5 Seaborn

Seaborn (Figura 10) é outra biblioteca muito utilizada para a construção de gráficos com o Python, que é baseada no Matplotlib e provê uma interface mais bem elaborada para construir gráficos com foco em estatística [22].

Figura 10. Seaborn Logo.

Fonte: [22].

2.3.6 Tweepy

Tweepy (Figura 11) é uma biblioteca para usar a API do Twitter em Python. Eficiente e de fácil utilização, disponibiliza diversos métodos que substituem as chamadas REST (*Representational State Transfer*) aceitas pela API do Twitter [23].

Figura 11. Tweepy logo.

Fonte: [23].

2.3.7 NLTK

NLTK (Figura 12) é uma biblioteca em Python para trabalhar com dados de linguagem humana. Ela fornece interfaces fáceis de usar para mais de 50 recursos, como WordNet, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização (processo de separação de um trecho de texto em unidades menores), lematização (processo de agrupar diferentes formas flexionadas de uma palavra para permitir sua analise como um único item), análise e raciocínio semântico, *wrappers* para bibliotecas PLN [25].

Figura 12. NLTK Logo.

Fonte: [25].

2.3.8 Jupyter

Jupyter é o mais recente ambiente de desenvolvimento interativo baseado em notebooks na web, código e dados. Sua interface flexível permite que os usuários configurem e organizem fluxos de trabalho em ciência de dados, computação científica e aprendizado de máquina. [33].

Figura 13. Jupyter.

Fonte: [33]

2.4 Análise de Sentimentos

A Análise de Sentimentos, também conhecida como Mineração de Opiniões é definida como a área de estudo que analisa opiniões, sentimentos, avaliações e atitude das pessoas em relação a um produto, serviço, organização, individuo, entre outros. A Análise de Sentimentos evidencia principalmente opiniões que expressam ou implicam sentimentos positivos ou negativos [24].

Essa área surge para extrair informações sobre textos utilizando técnicas de mineração de dados e processamento de linguagem natural. Os primeiros trabalhos sobre a área datam do ano de 2002 e de lá para cá, cada vez mais pessoas têm buscado por opiniões *online*.

Para obter bons resultados na análise sentimental, é necessário implementar técnicas anteriores à análise propriamente dita, afim de facilitar a procura por polaridade num texto. Caso o texto seja informal, será necessário um pré-processamento, de modo a corrigir erros ortográficos e de pontuação, que posteriormente iriam dificultar o procedimento de análise [5].

O processamento de linguagem natural pode envolver diversas fases. Algumas mais simples como a divisão do texto em termos mais simples (*tokenizer*), até mais complexos, como análise sintática (*phrase chunking*) [25].

As técnicas de análise podem ser agrupadas em dois tipos: técnicas simbólicas e de aprendizagem de máquina. Técnicas simbólicas são mais simples, caracterizando-se pela

avaliação das palavras recorrendo a recursos léxicos, determinando se o seu sentido é positivo ou negativo.

2.4.1 Definição de Opinião

Uma opinião é uma quádrupla (g, s, h, t) onde g é a opinião alvo, s é o sentimento em relação ao alvo, h é o dono da opinião e t é o tempo no qual a opinião é expressada. Porém essa é uma definição não muito utilizada na prática, devido ao fato que na maior parte de reviews e casos de interesse a descrição completa do alvo pode não estar na mesma sentença [6].

Para complementar essa opinião surge um quinto elemento, a entidade e. A entidade é um serviço, pessoa ou tópico descrito por um par e: (T,W), onde T é uma hierarquia de partes e subpartes e W é o conjunto dos atributos de e. Cada subparte também possui o próprio conjunto de atributos [6].

Podemos definir uma opinião como uma quíntupla [24] ($e_i, a_{ij}, s_{ijkl}, h_k, t_l$), onde e_i é o nome de uma entidade; a_{ij} é um aspecto de e_i; s_{ijkl} é o sentimento no aspecto a_{ij} da entidade e_i; h_k é o detentor da opinião e t_l é o tempo em que a opinião é expressa por h_k [5]. Neste trabalho, a entidade é autor do tweets, aspectos são as palavras do conteúdo do tweet. O objetivo é analisar o sentimento de cada aspecto, e obtendo o sentimento da tweet como um todo.

O sentimento s_{ijkl} pode ser positivo, negativo, neutro ou expressado por níveis de intensidade distintos. Pode-se observar o nível distinto de intensidade em pesquisas de opinião e questionários, onde, partindo de uma afirmação ou opinião existem diferentes repostas, como “discordo” ou “discordo plenamente” [5].

2.4.2 Dicionários Léxicos

Um dicionário léxico é um dicionário de palavras que associa uma palavra a um valor de sentimento positivo ou negativo. Existem três formas de construir um dicionário léxico, construção manual, métodos baseados em e dicionário e em *corpus* [31].

Nesse trabalho serão utilizados três dicionários léxicos:

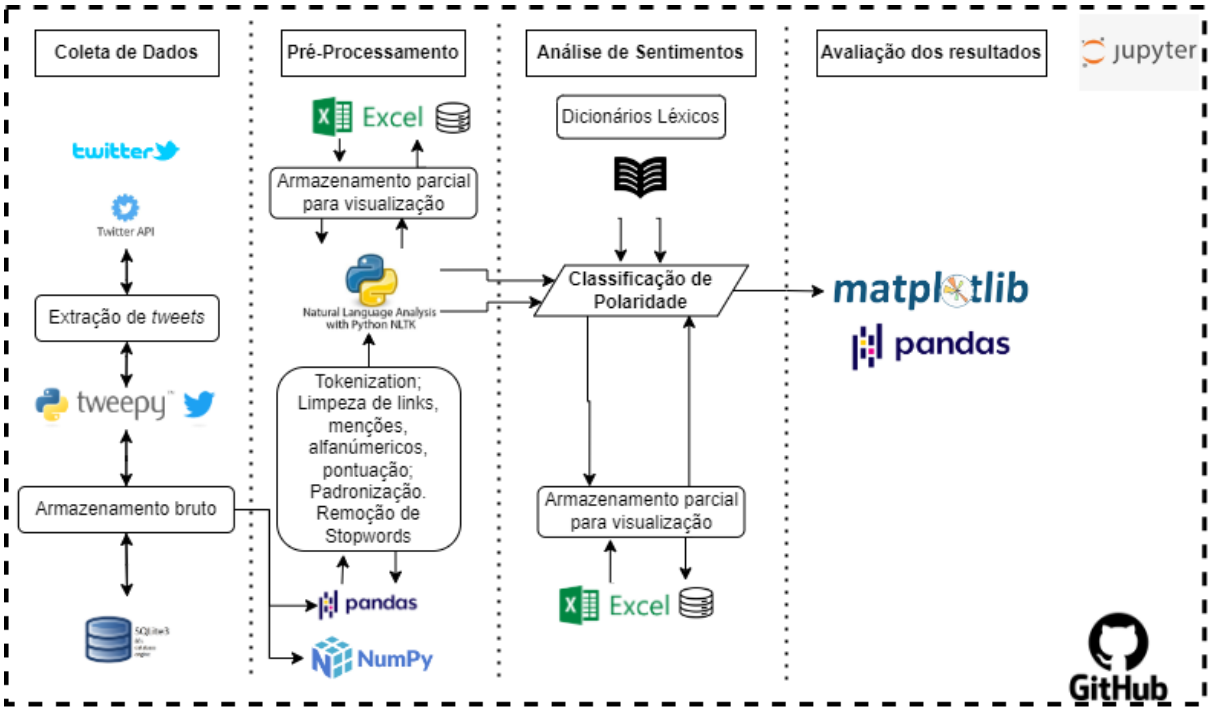
- *SentiWordNet*, um recurso léxico para mineração de opinião. Atribui cada a cada palavra do *WordNet* três pontuações de sentimento: positivo, negativo ou neutro.[27]
- OpLexicon, um dicionário léxico de sentimento para a língua portuguesa que contém palavras classificadas com a sua categoria morfológica e positividade positiva, negativa ou neutra. Criado e distribuído pelo Grupo de Processamento da Linguagem Natural da PUCRS [28].
- SentiLex é um dicionário léxico de sentimento concebido para a análise de sentimento sobre entidades humanas em textos redigidos em português. Atualmente é constituído por 7014 lemas e 82347 formas flexionadas [32].

3. Metodologia

Neste capítulo será apresentada a metodologia proposta para este trabalho, assim como os materiais produzidos durante o estudo, visando a aplicação dos conceitos teóricos vistos nos capítulos anteriores.

Na Figura 14 é ilustrada a metodologia proposta desse trabalho. As etapas seguidas foram: a coleta de dados; pré-processamento; análise de sentimentos; e, a avaliação dos resultados.

Figura 14. Metodologia proposta.



Fonte: Elaborada pelo autor.

3.1 Ferramentas

Todo este trabalho foi feito utilizando a linguagem de programação Python, utilizando a ferramenta Jupyter Notebook. O Jupyter Notebook é um aplicativo web open-source baseado em Python, que fornece um ambiente para experimentação de ideias, de uma forma parecida com cadernos de nota, permitindo escrever e guardar informações em tempo de execução.

Para a extração dos tweets através da API disponibilizada pelo próprio Twitter, foi utilizada a biblioteca *Tweepy*. Os dados brutos foram salvos em um banco de dados relacional SQLite3, onde cada *tweet* e suas informações era uma linha.

Para o pré-processamento, foram utilizadas: bibliotecas Pandas e Numpy para o gerenciamento e operação com os dados; biblioteca NLTK de Linguagem Natural e suas funções para tokenização e stopwords (palavras comumente usadas em uma língua, que não expressam sentimento).

Para a análise sentimental foram utilizados 3 dicionários léxicos em português diferentes:

LexiconPT [26], SentiWordNetPT-BR [27], e OpLexicon [28]. Para a avaliação dos resultados foram utilizadas as bibliotecas Matplotlib e Seaborn para a plotagem de gráficos.

Todos os materiais gerados estão disponíveis no GitHub do autor (https://github.com/carlospcneto).

3.2 A Coleta de Dados

Disponível em "notebook-coleta_de_dados.ipynb".

Após a importação das bibliotecas, é necessário fornecer as credenciais de autenticação para API do Twitter. Quando se deseja acessar a API é solicitado o registro de um aplicativo. Por padrão, aplicativos só podem acessar informações públicas no Twitter. Permissões não são dadas por padrão [29]. Neste trabalho foi utilizado a API v2 com nível de acesso *Essential*.

Depois de autenticado na API, monta-se a solicitação através da interface do *Tweepy*, utilizando a função "search_recent_tweets" (Figura 16). Tal função pode receber diversos parâmetros (Figura 15). Utilizou-se:

- *query*: recebe a/as palavras chaves;
- tweet_fields: campos que serão retornados em cada *tweet* pela api;
- max_results: limita a quantidade de tweets retornados pela função;
- start_time e end_time: definem o período dos tweets que serão retornados.

Para o armazenamento dos dados brutos é necessário conectar em um arquivo ".bd" com o auxílio da interface da biblioteca sqlite3 (estrutura banco de dados relacional). Após conectar-se, basta criar a tabela para armazenar os dados utiliza SQL.

Figura 15. Parâmetros da chamada para extração de tweets.

```
# Lula na CNN, 13/09/2022 20hrs
start_time = '2022-09-13T20:00:00Z'
end_time = '2022-09-14T04:00:00Z'
query = 'Lula'
```

Fonte: Elaborada pelo autor.

Figura 16. Chamada via Client na biblioteca Tweepy.

```
response = client.search_recent_tweets(query=query,
                    tweet_fields=['author_id','created_at', 'lang', 'text'],
                    max_results=100,
                    start_time=start_time, end_time=end_time)
```

Fonte: Elaborada pelo autor.

Depois de repetir essa chamada, basta usar do SQL para inserir os dados na tabela desejada. Do período de 6 horas após o início da entrevista de interesse, foram salvos 199655 *tweets* no banco de dados.

3.3 Pré-processamento

Como é possível ver na Figura 17, os *tweets* brutos tem diversos problemas em seu texto. Nesta etapa, disponível no notebook "notebook-pre_processamento.ipynb", trata-se o texto da coluna "text_tweet" de forma a deixar o texto pronto para a análise.

Figura 17. Base não formatada.

Table: tweets_cnn — Filter in any column

	id_tweet	text_tweet	createdat_tweet	lang_tweet	id_author
	Filter	Filter	Filter	Filter	Filter
1	1569898726503620608	RT @CMonteroOficial: Se viene una paliza en ...	2022-09-14T03:59:58.000Z	es	833485287363862529
2	1569898725585166338	RT @SRivoltril: O mimimizento já se recuperou ...	2022-09-14T03:59:58.000Z	pt	210483399
3	1569898725115416576	@Janoninho_ Sou Lula ❤ ⭐BR	2022-09-14T03:59:58.000Z	pt	227177141
4	1569898724956192769	@ThaiscHackbart @johnny22purin ...	2022-09-14T03:59:57.000Z	pt	1288310849950670848
5	1569898724133937154	RT @kimpaim: Perfil de coluna da GLOBO está ...	2022-09-14T03:59:57.000Z	pt	50512358
6	1569898722510651397	RT @ViniciusPoit: A verdade tem que ser dita:...	2022-09-14T03:59:57.000Z	pt	100072067
7	1569898721369899009	RT @ptsaopaulosp: Lula Presidente e Haddad ...	2022-09-14T03:59:57.000Z	pt	33582597
8	1569898720917073920	RT @ImpuestosyE: Lula da Silva aumenta a 15 ...	2022-09-14T03:59:57.000Z	es	1042916961921843200
9	1569898719608176640	RT @PedroRonchi2: Lula criou o Samu, Bolsonar...	2022-09-14T03:59:56.000Z	pt	3092078535
10	1569898719126102018	RT @DiarideGirona: Científics de la UdG ...	2022-09-14T03:59:56.000Z	ca	850271746464768000
11	1569898717225881600	RT @senadorhumberto: A Senadora de Lula é ...	2022-09-14T03:59:56.000Z	pt	1536813610860691458
12	1569898716500295681	RT @antissiemssia: A ministra Rosa Weber, que ...	2022-09-14T03:59:55.000Z	pt	1492998675668340744
13	1569898714470223874	RT @musaraujo: gente plmdds se Lula ganhar n...	2022-09-14T03:59:55.000Z	pt	1107370210338709504
14	1569898709072084998	RT @lulafalcao: Em editorial, o The Guardian to...	2022-09-14T03:59:54.000Z	pt	330749067
15	1569898704991145985	RT @ritaduF: A impressão que tenho é que os ...	2022-09-14T03:59:53.000Z	pt	1374544572320190464
16	1569898703137259520	RT @siteptbr: Lula na CNN: "Nós tínhamos o ...	2022-09-14T03:59:52.000Z	pt	1389081163403337728
17	1569898702554255360	RT @SigaGazetaBR: Bolsonaro: "imprensa teve ...	2022-09-14T03:59:52.000Z	pt	356552458
18	1569898696795471872	@SamPancher @Metropoles https://t.co/...	2022-09-14T03:59:51.000Z	qme	1567952399830028289

Fonte: Elaborada pelo autor.

Após carregar os tweets do Banco de Dados, foi executada uma conversão para um DataFrame do pandas, já filtrando os *tweets* apenas em português, de forma a facilitar o manuseio dos dados.

Figura 18. Transformação da base de dados para DataFrame.

```
In [9]: # Transformando em um DataFrame
        bd_array = np.array(data_bd)
        df = pd.DataFrame(bd_array, columns = ['id_tweet','text_tweet','createdat_tweet'
        df = df[(df['lang_tweet'] == 'pt' )]
        print(df.head())
```

```
              id_tweet                                         text_tweet  \
0  1569777944377294848  @jjneto09 @monicabergamo VAMOS CUIDAR QUE ESSE...
1  1569777944033206278  RT @leiatheinvestor: ++ Assessor econômico de ...
2  1569777943999561737  RT @cartacapital: ÀS 16H | Lula tem 46% no IPE...
3  1569777943966089216  RT @PedroRonchi2: Quando o Lula voltar a gover...
4  1569777943626149888  Aliados de Bolsonaro exploram nas redes sociai...

           createdat_tweet lang_tweet           id_author
0  2022-09-13T20:00:01.000Z         pt  1512063959251120142
1  2022-09-13T20:00:01.000Z         pt   812283375637721088
2  2022-09-13T20:00:01.000Z         pt           1024584565
3  2022-09-13T20:00:01.000Z         pt            230723137
4  2022-09-13T20:00:01.000Z         pt              9317502
```

Fonte: Elaborada pelo autor.

Após esse primeiro filtro, são definidos alguns outros processamentos para o texto:

- Remover linhas duplicadas;
- Remover parte originaria do *Retweet*;
- Remover menções a usuários;
- Remover links;
- Remover numéricos e caracteres especiais;
- Remover pontuações;
- Remover quebras de linha;
- Remover espaços duplos;
- Remover a palavra utilizada na query de busca (“lula”);
- Padronizar todo o texto para letras minúsculas;
- Remover letras duplas (exceto, ‘ss’ e ‘rr’).

A Figura 19 apresenta os 5 primeiros *tweets* do DataFrame após o pré-processamento na coluna “processed_text”. A remoção de linhas duplicadas (comumente *retweets* ou “copia e cola” de texto) reduziu a quantidade de *tweets* de 199655 para 52258, aproximadamente 26,2% do volume original.

Figura 19. Resultado do pré-processamento.

In [19]:
```
# Remove linhas duplicadas
df.drop_duplicates(['text_tweet'], inplace=True)
print(len(df))
df['processed_text'] = df['text_tweet'].map(clean_tweet)
df.head()
```

```
52258
```

Out[19]:

id_tweet	text_tweet	createdat_tweet	lang_tweet	id_author	processed_text
7944377294848	@jjneto09 @monicabergamo VAMOS CUIDAR QUE ESSE...	2022-09-13T20:00:01.000Z	pt	1512063959251120142	vamos cuidar que esse desespero deles não se ...
7944033206278	RT @leiatheinvestor: ++ Assessor econômico de ...	2022-09-13T20:00:01.000Z	pt	812283375637721088	assessor econômico de lula disse que eventual...
7943999561737	RT @cartacapital: ÀS 16H	Lula tem 46% no IPE...	2022-09-13T20:00:01.000Z	pt	1024584565	às lula tem no ipec o que falta para definir ...
7943966089216	RT @PedroRonchi2: Quando o Lula voltar a gover...	2022-09-13T20:00:01.000Z	pt	230723137	quando o lula voltar a governar os bolsonaris...
7943626149888	Aliados de Bolsonaro exploram nas redes sociai...	2022-09-13T20:00:01.000Z	pt	9317502	aliados de bolsonaro exploram nas redes sociai...

Fonte: Elaborada pelo autor.

3.4 Análise de Sentimentos

Esta etapa está disponível no notebook "notebook-analise_sentimental.ipynb". Nesta etapa que acontece de fato a análise sobre os *tweets* coletados e pré-processados. A análise ocorrerá a nível de sentença em duas etapas: a *tokenização* e remoção de *stopwords*, momento onde os *tweets* terão suas palavras separadas, e a classificação, onde cada palavra terá um score (que variam de -1 a 1) de acordo com o score fornecido pelo dicionário utilizado.

Essa pontuação (Figura 20) será totalizada para cada tweet, assim teremos um resultado de polaridade para cada tweet de acordo com o dicionário.

Figura 20. Classificação final para cada tweet.

```
In [12]: df_final.head()
```

Out[12]:

	lang_tweet	id_author	processed_text	pos_processed_text	polaridade	score	dicionario
	pt	1512063959251120142	vamos cuidar que esse desespero deles não se ...	[vamos, cuidar, desespero, torne, causa, fraud...	Negativo	-1.875	SentiWord Pt-BR v1.0b
	pt	9317502	aliados de bolsonaro exploram nas redes sociai...	[aliados, bolsonaro, exploram, redes, sociais,...	Neutro	0.0	lexico_v3.0
	pt	230723137	quando o voltar a governar os bolsonaristas v...	[voltar, governar, bolsonaristas, vão, enxerga...	Neutro	0.0	lexico_v3.0
	pt	1024584565	às tem no ipec o que falta para definir a vit...	[ipec, falta, definir, vitória, turno, assista...	Neutro	0.0	lexico_v3.0
	pt	812283375637721088	assessor econômico de disse que eventual gove...	[assessor, econômico, disse, eventual, governo...	Neutro	0.0	lexico_v3.0

Fonte: Elaborado pelo autor.

4. Resultados e Discussões

Neste capítulo, os resultados obtidos no trabalho, disponíveis em "notebook-resultados.ipynb", serão apresentados, discutidos e comparados entre si e com outro trabalhos. Como resultado direto da classificação usando os dicionários léxicos, obteve-se em números absolutos (Tabela *1* e Figura 21) e em porcentagem (Tabela *2*), lembrando que o número total de *tweets* classificados é de 52258.

Tabela 1. Resultados classificação polaridade por dicionário em quantidade.

Sentimento / Dicionário	**SentiLex-lem-PT01**	**SentiWord Pt-BR v1.0b**	**lexico_v3.0**	**Média**
Positivo	7140	16416	11613	11723
Neutro	31888	19031	24959	25292
Negativo	13230	16811	15686	15242

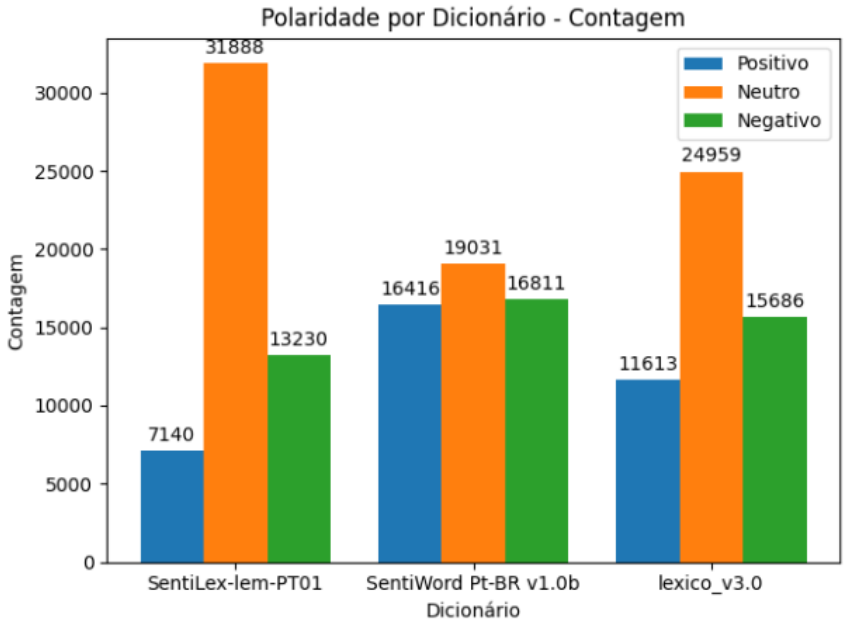
Fonte: Elaborado pelo autor.

Tabela 2. Resultados classificação polaridade por dicionário em porcentagem.

Sentimento / Dicionário	**SentiLex-lem-PT01**	**SentiWord Pt-BR v1.0b**	**lexico_v3.0**	**Média**
Positivo	14%	31%	22%	22%
Neutro	61%	36%	48%	49%
Negativo	25%	32%	30%	29%

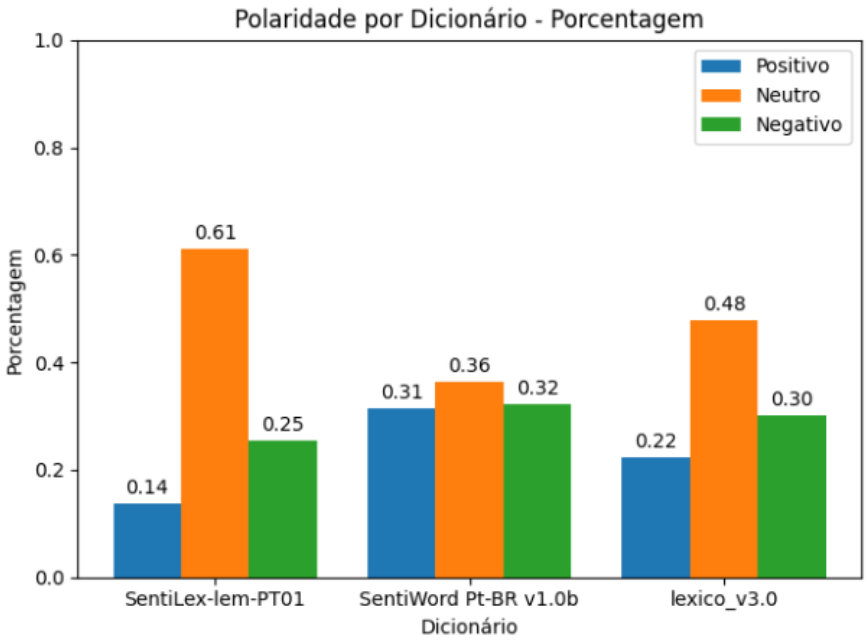
Fonte: Elaborado pelo autor.

Figura 21. Gráfico polaridade por dicionário – contagem.



Fonte: Elaborado pelo autor.

Figura 22. Gráfico polaridade por dicionário – porcentagem.



Fonte: Elaborado pelo autor.

Inicialmente deve-se notar a alta proporção de tweets classificados como neutros em todos os dicionários. Associa-se esses fatos a alguns pontos importantes:

- O uso nos *tweets* de muitos termos que acabam não impactando no score, como nomes, comentários sobre a política em site, entre outros.
- Muitos *tweets* não expressam o sentimento de forma clara e objetiva, dificultando a análise do real sentimento.

Essa alta neutralidade deve ser considerada e estudada em estudos futuros. Em relação a proporção de *tweets* classificados como positivos e neutros, vemos que a predominância é *tweets* com polaridade negativa.

Analisando os tweets extraídos é possível compreender que independente do candidato político, a maioria dos usuários utiliza-se de termos para ataques, termos pejorativos nos seus *tweets* durante períodos em que um candidato à presidência esteja em "alta", conforme visto na , um *tweet* que diz sobre um aumento no apoio a Lula e outro "reclamando" da entrevista são classificados como negativo em um dicionário e neutro nos outros, conforme Tabela 3.

Tabela 3. Exemplos de tweets.

Texto do *tweet*	**Classificação lexico_v3.0**	**Classificação SentiLex-lem-PT01**	**Classificação SentiWord Pt-BR v1.0b**
"A BARRAGEM ROMPEU: IPEC DETECTA FORTE MIGRAÇÃO DE CIRISTAS PARA LULA", de um ex-cirista, Miguel	0.0	0.0	-0.75
"Ain ele tinha registro CAC". Então, mídia de MERDA, E ESSA TATUAGEM DO LULA????? Mídia lixo. @folha lixo. Quadrilha militante do PT. Não passam disso. https://t.co/a1nHABa8YY	0.0	0.0	-1.0

Nos outros trabalhos outrora aqui citados, relatou-se também a dificuldade da análise via dicionários léxicos. COSTA 2017, relatou também as limitações dos dicionários léxicos, porém seu trabalho teve êxito em demonstrar o uso das ferramentas e metodologia apresentadas. MONTELES 2019, que também utilizou 3 dicionários, executou melhorias nos dicionários de forma a aumentar a acurácia em comparação a uma avaliação manual dos *tweets*.

5. Conclusões

Este trabalho teve como objetivo mostrar o processo de análise sentimental com diferentes dicionários à nível de sentença e extração de dados usando da linguagem Python. Na análise, de mais de 50 mil tweets durante e após a entrevista do candidato à presidência Lula na CNN Brasil, no dia 13/09 de 2022, foram obtidos resultados de polaridade para diferentes dicionários léxicos de aproximadamente 29% de sentenças classificadas como negativa, 22% como positiva e 49% neutro.

Dentro do que se abrange as ferramentas e metodologia utilizados, o trabalho obteve êxito em questões importantes no processamento, como uso de memória e tempo de extração e processamento.

Após os resultados, constatou-se que a análise de sentimentos a nível de sentença não é uma tarefa fácil, pois não existe muita clareza de opinião na maioria dos tweets.

Para trabalhos futuros será necessário utilizar de outras formas de análise sentimental, como a classificação em pares, ao invés de palavras individuais.

6. Referências

[1] TIC Domicílios 2021, Lançamento dos Resultados. Cetic.br, São Paulo, 2022. Disponível em: <https://cetic.br/media/analises/tic_domicilios_2021_coletiva_imprensa.pdf>. Acesso em: setembro de 2022.

[2] Wright. A. Análise de sentimentos é o novo campo na web. Tecnlogia Terra, 25 de agosto de 2009. Disponível em: <http://tecnologia.terra.com.br/internet/analise-de-sentimentos-e-novo-campo-na-web,48e8887dc5aea310VgnCLD200000bbcceb0aRCRD.html>. Acesso em: setembro de 2022.

[3] Tsytsarau M. and Palpanas T. (2010). Survey on Mining Subjective Data on the Web. In Proceedings of the 3 rd workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Data Mining and Knowledge Discovery.

[4] Baptista. R. Redes sociais influenciam voto de 45% da população, indica pesquisa do DataSenado. Disponível em: <https://www12.senado.leg.br/noticias/materias/2019/12/12/redes-sociais-influenciam-voto-de-45-da-populacao-indica-pesquisa-do-datasenado>. Acesso em: setembro de 2022.

[5] MONTELES, G. S. ANÁLISE DE SENTIMENTO: uma comparação de dados extraídos do Twitter a partir de diferentes dicionários léxicos. Universidade Federal do Maranhão, 2019.

[6] COSTA, A. C. B. Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o framework Apache Ignite. Universidade Federal do Maranhão, 2017.

[7] Braun, D. Brasil tem a quarta maior base de usuários do Twitter no mundo. Globo. 2022 Disponível em: <https://valorinveste.globo.com/mercados/internacional-e-commodities/noticia/2022/04/25/brasil-tem-a-quarta-maior-base-de-usuarios-do-twitter-no-mundo.ghtml>. Acesso em: setembro de 2022.

[8] Brasil é o 3º país que mais usa redes sociais no mundo: 1º- Youtube e 2º- WhatsApp. Disponível em: <https://www.diariozonanorte.com.br/brasil-e-o-3o-pais-que-mais-usa-redes-sociais-no-mundo-1o-youtube-e-2o-whatsapp/>. Acesso em: setembro de 2022.

[9] TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, v. 8, p. 53-65, 2011.

[10] CHEN, M. et al. Big Data Related Technologies, Challenges and Future Prospects. SpringerBriefs in Computer Science, 2014.

[11] Apache Hadoop. Disponível em: <https://hadoop.apache.org/>. Acesso em: setembro de 2022.

[12] O que é Big Data? Oracle. Disponível em: <https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em: setembro de 2022.

[13] Helder. Os 5Vs do Big Data. Disponível em: <https://culturaanalitica.com.br/os-5-vs-big-data/>. Acesso em: setembro de 2022.

[14] GADOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. 2015.

[15] CAVANILLAS, J. M. et al. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. [S.l.]: SpringerBriefs in Computer Science, 2016.

[16] Python. Disponível em: <https://www.python.org/about/>. Acesso em: setembro de 2022.

[17] PySpark Documentation. Disponível em: <https://spark.apache.org/docs/latest/api/python/>. Acesso em: setembro de 2022.

[18] Numpy, The fundamental package for scientific computing with Python. Disponível em: <https://numpy.org/> Acesso em: setembro de 2022.

[19] Pandas Documentation. Disponível em: <https://pandas.pydata.org/docs/index.html/> Acesso em: setembro de 2022.

[20] Scikit Learn. Disponível em: <https://scikit-learn.org/stable/about.html#citing-scikit-learn/> Acesso em: setembro de 2022.

[21] Matplotlib. Disponível em: <https://matplotlib.org/> Acesso em: setembro de 2022.

[22] Seaborn. Disponível em: <https://seaborn.pydata.org/citing.html> Acesso em: setembro de 2022.

[23] Tweepy API. Disponível em: <https://www.tweepy.org/> Acesso em: setembro de 2022.

[24] LIU, B. Sentiment Analysis and Opinion Mining. Chicago: Morgan & Claypool Publishers, 2012. 168 p.

[25] NLTK Docs. Disponível em: <https://www.nltk.org/> Acesso em: setembro de 2022.

[26] LexiconPT. Disponível em: <https://github.com/sillasgonzaga/lexiconPT> Acesso em: setembro de 2022.

[27] SentiWordNet. Disponível em: <https://github.com/Pedro-Thales/SentiWordNet-PT-BR> Acesso em: setembro de 2022.

[28] OPLEXICON. Disponível em: <https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/> Acesso em: setembro de 2022.

[29] Twitter for Developers. Disponível em: <https://developer.twitter.com/en/portal/products/essential> Acesso em: setembro de 2022.

[30] O que é NoSQL?. Disponível em: <https://www.oracle.com/br/database/nosql/what-is-nosql/> Acesso em: setembro de 2022.

[31] Jurek A., Mulvenna M. D. Improved lexicon-based sentiment analysis for social media analytics, 2015.

[32] Carvalho P., Silva M. J. Sentilex-pt: principais características e potencialidades. 2015. Journals, Oslo Studies in Language.

[33] JupyterLab: A Next-Generation Notebook Interface. Disponível em: <https://jupyter.org/> Acesso em: setembro de 2022.