MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Especialização em Ciência de Dados

Monitoramento de Dados para Suporte ao Relacionamento com Clientes: Proposição de uma Ferramenta de Busca e Processamento de Informações para Desenvolvimento de *Insights*.

Cleuber Lúcio da Silva Rodrigues

João Monlevade, MG
2024

Cleuber Lúcio da Silva Rodrigues

Monitoramento de Dados para Suporte ao Relacionamento com Clientes: Proposição de uma Ferramenta de Busca e Processamento de Informações para Desenvolvimento de *Insights*.

Trabalho de conclusão de curso apresentado ao curso de Ciência de Dados do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

R696m Rodrigues, Cleuber Lucio da Silva.
Monitoramento de dados para suporte ao relacionamento com clientes [manuscrito]: proposição de uma ferramenta de busca e processamento de informações para desenvolvimento de insights. / Cleuber Lucio da Silva Rodrigues. - 2024.
24 f.

Orientador: Prof. Dr. thiago augusto de oliveira Silva.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Processamento de linguagem natural (Computação). 2. Processamento de textos (Computação). 3. Processamento eletrônico de dados - Sites da Web. 4. Processo decisório. 5. Proteção de dados. 6. Sistemas de coleta automática de dados. I. Silva, thiago augusto de oliveira. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8:004.6

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO - ICEA

FOLHA DE APROVAÇÃO

Cleuber Lúcio da Silva Rodrigues

Monitoramento de Dados para Suporte ao Relacionamento com Clientes: Proposição de Uma Ferramenta de Busca e Processamento de Informações para Desenvolvimento de *Insights*

Trabalho de conclusão de curso apresentado ao curso de Especialização em Ciência de Dados da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Especialista em Ciência de Dados

Aprovada em 03 de julho de 2024

Membros da banca

Dr. Thiago Augusto de Oliveira Silva - Orientador - Universidade Federal de Ouro Preto
Dr. Sergio Evangelista Silva - Universidade Federal de Ouro Preto
Me. Ronaldo Neves Ribeiro - Celulose Nipo Brasileira - CENIBRA

Thiago Augusto de Oliveira Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 29/07/2024

Documento assinado eletronicamente por **Thiago Augusto de Oliveira Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 01/08/2024, às 17:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0750941** e o código CRC **27C29B36**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0817 - www.ufop.br

Ao meu pai, meu guia e exemplo de perseverança. Sua sabedoria e seus exemplos são as âncoras que me sustentam em todas as tempestades. Este trabalho é dedicado a você, com imensa gratidão e admiração.

Ao meu querido filho, que ilumina meus dias com seu sorriso e renova minha esperança no futuro. Que este trabalho sirva como um testemunho do meu compromisso em construir um mundo melhor para você e as gerações futuras.

À minha esposa, minha companheira de jornada, cujo amor e apoio incondicional são meu refúgio seguro. Sua presença ao meu lado torna todas as conquistas mais significativas.

Agradecimentos

À Deus pelas oportunidades que tive na vida.

Aos meus pais pelo amor incondicional, por me mostrarem o significados destas palavras e por fomentarem meus sonhos.

À minha família por me apoiar em todos os momentos.

À CENIBRA pelas oportunidades de desenvolvimento pessoal e profissional.

Ao Adermo-San pela confiança, pela amizade e pelos bons momentos de discussão acerca do trabalho.

Aos amigos, Luiz, Gustavo e Igor pela amizade e pela imensa contribuição ao trabalho.

Ao Prof. Thiago pela orientação, incentivo e confiança.

Aos amigos do DECOM e DC pela convivência diária.

Àqueles que direta ou indiretamente contribuiram na realização desse trabalho meus sinceros agradecimentos!

“A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.”

Albert Einstein

Resumo

Este trabalho investiga a relevância do monitoramento de dados para embasar processos decisórios e fortalecer o relacionamento com clientes, propondo a concepção de uma ferramenta dedicada à busca e processamento de informações para a geração de *insights*. Destaca-se o desafio representado pela abundância de dados e a complexidade inerente à análise efetiva de dados textuais, provenientes de diversas fontes voltadas a uma área específica do conhecimento. São abordadas as técnicas de *web crawling* e *web scraping*, juntamente com a necessidade de compreender as leis de proteção de dados. A proposta concentra-se na criação de uma ferramenta destinada à busca e estruturação de informações pertinentes ao setor de celulose e papel, visando oferecer *insights* sobre produtos e mercados para facilitar a tomada de decisões. O estudo analisa as implicações legais e éticas do *web scraping* e *crawling*, enfatizando a importância de aderir às práticas recomendadas para mitigar potenciais impactos negativos nos *sites* fonte. Descreve-se a metodologia de desenvolvimento de um algoritmo em *Python* para a pesquisa de notícias baseada em palavras-chave, destacando a definição de objetivos e escopo da ferramenta, fontes de dados, considerações legais e o processo de desenvolvimento do algoritmo. Além disso, são discutidos aspectos como a indexação de dados em formato JSON e a documentação do código, processos e decisões de *design*. Os resultados obtidos, exemplificados pelo uso da palavra-chave "celulose"na ferramenta, incluem a recuperação de informações, análise de sentimentos e estruturação de dados, demonstrando eficiência na obtenção e análise de notícias relacionadas ao setor de celulose e papel. A solução evidenciou resultados positivos na análise de sentimentos, revelando uma tendência predominantemente positiva na cobertura de notícias sobre o setor. A utilização de algoritmos de *machine learning* e processamento de linguagem natural permitiu a classificação de notícias em diferentes categorias de sentimentos, contribuindo para a compreensão da percepção pública e para o monitoramento da reputação do setor. Conclui-se que a aplicação de técnicas avançadas, como *web crawling*, *web scraping* e análise de sentimentos, possibilitou a recuperação e análise de dados de forma eficaz, oferecendo contribuições significativas para a compreensão do cenário e para a tomada de decisão informada no setor.

Palavras-chaves: Ferramenta de busca. *Crawling*. *Scraping*. Processamento de linguagem natural. *Word Cloud*.

Abstract

This work investigates the relevance of data monitoring to support decision-making processes and strengthen customer relationships, proposing the conception of a tool dedicated to searching and processing information for generating insights. The challenge represented by the abundance of data and the inherent complexity of effectively analyzing textual data from various sources focused on a specific area of knowledge is highlighted. Techniques such as web crawling and web scraping are addressed, along with the need to understand data protection laws. The proposal focuses on creating a tool for searching and structuring information relevant to the pulp and paper sector, aiming to provide insights into products and markets to facilitate decision-making. The study examines the legal and ethical implications of web scraping and crawling, emphasizing the importance of adhering to recommended practices to mitigate potential negative impacts on source websites. The methodology for developing a Python algorithm for keyword-based news search is described, highlighting the definition of objectives and scope of the tool, data sources, legal considerations, and the algorithm development process. Additionally, aspects such as data indexing in JSON format and code documentation, processes, and design decisions are discussed. The results obtained, exemplified by the use of the keyword "pulp" in the tool, include information retrieval, sentiment analysis, and data structuring, demonstrating efficiency in obtaining and analyzing news related to the pulp and paper sector. The solution showed positive results in sentiment analysis, revealing a predominantly positive trend in news coverage of the sector. The use of machine learning algorithms and natural language processing enabled the classification of news into different sentiment categories, contributing to the understanding of public perception and sector reputation monitoring. It was concluded that the application of advanced techniques such as web crawling, web scraping, and sentiment analysis allowed for the effective retrieval and analysis of data, offering significant contributions to understanding the landscape and making informed decisions in the sector.

Keywords: Searching tool. Crawling. Scraping. Sentiment analysis. Word Cloud.

Lista de ilustrações

Lista de Abreviaturas e Siglas

API Application Programming Interface

CFAA Computer Fraud and Abuse Act

CRM Customer Relationship Management

CSS Cascading Style Sheets

DOM Document Object Model

HTML HyperText Markup Language

HTTP Hypertext Transfer Protocol

IC Inteligência Competitiva

JSON JavaScript Object Notation

NLTK Natural Language Toolkit

PLN Processamento de Linguagem Natural

URL Uniform Resource Locator

Sumário

1 Introdução

As empresas contemporâneas enfrentam o desafio incessante de gerar vastas quantidades de dados, sejam relacionados ao desempenho operacional ou provenientes de notícias. Contudo, essa proliferação de informações nem sempre se traduz em valor tangível (CHEN *et al.*, 2012). Embora seja reconhecido que dados representam ativos valiosos para a tomada de decisão, a análise de grandes volumes de informações ainda não se incorporou totalmente às estratégias de muitos gestores (MCAFEE; BRYNJOLFSSON, 2012).

A complexidade advinda da presença de dados em diversos formatos e provenientes de múltiplas fontes emerge como um desafio significativo para uma análise eficaz (DAVENPORT; HARRIS, 2007). Os dados textuais chamam a atenção das organizações, pois milhares de pessoas se expressam diariamente em textos, como a descrição das percepções e comportamentos dos clientes bem como relatos de especialistas de mercado.

Os dados textuais dos clientes desempenham um papel crucial na tomada de decisões, embora sua utilização ainda seja modesta, principalmente devido às dificuldades na análise e interpretação desses dados provenientes de diversas fontes. Essa limitação na extração de *insights* significativos tem sido um desafio reconhecido na indústria (AWAN *et al.*, 2021).

A coleta de dados disponibilizados em sites públicos, em especial através de técnicas de *web crawling* e *web scraping*, possibilita alimentar bases de dados para diversas finalidades, permitindo o cruzamento desses dados para inferir informações adicionais sobre empresas, órgãos públicos, entidades ou indivíduos (AWAN *et al.*, 2021). Ao utilizar essas técnicas, os usuários devem estar atentos às implicações legais e éticas envolvidas. É essencial compreender e cumprir as leis de proteção de dados em vigor, buscando sempre a transparência, o consentimento informado e o uso responsável dos dados coletados (ALDEEN; SALLEH; RAZZAQUE, 2015).

A diferença entre "*web scraping*" e "*web crawling*" é relativamente vaga, já que muitos autores e programadores usarão ambos os termos de forma intercambiável. Em termos gerais, o termo "*crawler*" indica a capacidade de um programa de navegar por páginas da web por conta própria, talvez até sem um objetivo ou propósito final bem definido, explorando infinitamente o que um site ou a web tem a oferecer. Os "*Web crawlers*"são muito usados por mecanismos de pesquisa como o *Google* para recuperar o conteúdo de um URL, examinar essa página em busca de outros *links*, recuperar os URLs desses *links* e assim por diante (BROUCKE; BAESENS, 2018).

Web scraping é uma técnica para converter dados da *Web* não estruturados em dados estruturados que podem ser salvos e analisados em uma planilha ou banco de dados central. Isso permite que um programa recupere grandes volumes de dados em um curto espaço de tempo, o que é vantajoso no mundo de hoje, especialmente porque temos dados sempre mudando e sendo atualizados (THOMAS; MATHUR, 2019).

Ao usar das técnicas *web crawling* e *web scraping*, a empresa pode coletar um grande conjunto de dados das descrições dos produtos nos principais mercados mundiais, comentários de clientes e varejistas, além de *feedback* em redes sociais ou blogs dos clientes de seus clientes. A análise desses dados pode ajudar a área comercial a fornecer ao fabricante descrições mais claras de seus produtos em cada mercado, além de listar os problemas enfrentados pelos usuários finais com seus produtos e usar seus comentários para aprimoramento dos produtos e garantir sua lucratividade por meio das vendas (WANG; LAI; LIN, 2023).

1.1 Objetivo

O propósito do presente trabalho foi criar uma ferramenta para busca e indexação de assuntos vinculados ao setor de celulose e papel que possam fornecer *insights* sobre os produtos e principais mercados de atuação para auxílio no processo de tomada de decisão.

1.1.1 Objetivos específicos

- Recuperação de informações: Aplicar técnicas de scraping e crawler para recuperação de informações relevantes para a industria de celulose e papel.
- Estruturação da base de dados: Criar base de dados estruturada das informações recuperadas.
- Análises de sentimento: Aplicar metodologias de Processamento de Linguagem Natural para condicionar análises de sentimentos dos dados extraídos e estruturados;
- Nuvem de palavras: Aplicar metodologias de Processamento de Linguagem Natural para criação de nuvem de palavras para os dados extraídos e estruturados;

2 Revisão da Literatura

A indústria de celulose e papel é caracterizada por sua dinâmica complexa, que demanda acesso rápido e eficiente a dados relevantes do mercado para a tomada de decisões estratégicas. A indústria global de celulose e papel é altamente influenciada por flutuações no mercado global de commodities, incluindo mudanças nos preços de matérias-primas e insumos, demanda dos consumidores e condições geopolíticas e econômicas globais. A complexidade da cadeia de suprimentos, como destacado em relatório da Indústria Brasileira de Árvores (2022), também contribui para a dinâmica desafiadora enfrentada pela indústria.

Além disso, a indústria está sujeita a regulamentações ambientais globais rigorosas, que podem impactar significativamente as operações e os processos de produção das empresas. Diante desse cenário, as empresas do setor precisam buscar constantemente inovações em tecnologia, sustentabilidade e eficiência operacional para se manterem à frente da concorrência.

A proposição de uma ferramenta de *web scraping* e *crawling* em *Python*, com uso dos frameworks BeautifulSoup, Scrapy ou Requests-html, aliada ao Processamento de Linguagem Natural (PLN), pode oferecer uma abordagem poderosa para análise de dados (THOMAS; MATHUR, 2019). Ao integrar essas técnicas avançadas, os usuários relacionados ao setor podem obter *insights* valiosos para tomada de decisões estratégicas, melhorando sua competitividade e eficácia no mercado.

O BeautifulSoup e o Scrapy são dois dos *frameworks* mais populares em *Python* para *web scraping* e *crawling*. O BeautifulSoup oferece uma interface simples e intuitiva para extração de dados de páginas da *web* devido à sua simplicidade de uso e à clareza de sua sintaxe. Ele é particularmente eficaz para tarefas de *web* scraping mais simples, onde o objetivo é extrair informações de páginas *web* estáticas de maneira direta e sem a necessidade de configurações complexas (RICHARDSON, 2021).

Por outro lado, o *Scrapy* é considerado uma biblioteca mais avançada devido à sua capacidade de lidar com tarefas mais complexas de *web crawling* e *scraping*. Ele fornece recursos poderosos, como suporte para navegação *web* dinâmica, gerenciamento de sessões, identificação e extração de dados de maneira mais eficiente em sites complexos e estruturados (Scrapy Developers, acessado em 2024). Além disso, o Scrapy é altamente escalável, o que significa que pode lidar com grandes volumes de dados e processos de coleta em larga escala de maneira eficiente. Conforme Thomas e Mathur (2019), a combinação desses tipos de *frameworks* permite a coleta eficiente de dados de várias fontes *online*.

A biblioteca Requests-html em *Python* é também outra ferramenta para análise de páginas da *web* e *scraping*. Com suas funcionalidades abrangentes, ela oferece aos desenvolvedores uma maneira eficaz de fazer solicitações HTTP, analisar o conteúdo HTML de páginas da *web* e extrair dados relevantes (REITZ, 2019).

Segundo Reitz (2019), uma das principais características da requests-html é sua capacidade de fazer solicitações HTTP para páginas da *web*, permitindo que os desenvolvedores obtenham o conteúdo de páginas da *web* de forma rápida e eficiente. Além disso, a biblioteca oferece uma interface intuitiva para analisar o HTML das páginas da *web*, permitindo que os desenvolvedores naveguem pelo DOM da página e acessem elementos HTML. Uma das funcionalidades mais úteis da biblioteca Requests-html é seu suporte a seletores CSS e XPath. Isso permite que os desenvolvedores selecionem e extraiam elementos específicos de uma página da *web* com base em seus atributos, classes, identificadores e outros critérios de seleção.

Outra característica importante da biblioteca requests-html é sua capacidade de renderizar JavaScript em páginas da *web*. Isso é especialmente útil para páginas que usam JavaScript para renderizar conteúdo dinâmico, pois permite que os desenvolvedores acessem todo o conteúdo da página, incluindo elementos carregados dinamicamente. Além disso, a biblioteca requests-html oferece suporte a sessões persistentes e cookies, permitindo que os desenvolvedores mantenham o estado da sessão ao fazer várias solicitações HTTP para a mesma página da *web* (REITZ, 2019).

O PLN é uma área da ciência da computação que se concentra na interação entre computadores e linguagem humana. No contexto da análise de mercados, técnicas de PLN podem ser aplicadas para extrair *insights* significativos de grandes volumes de dados textuais. Isso inclui a identificação de tendências do mercado, análise de sentimento em comentários de clientes, consultores e analistas, bem como extração de informações de relatórios e notícias, entre outras aplicações (ARJUNAN, 2022).

No contexto da gestão de informações, o PLN pode ser aplicado de diversas maneiras para extrair *insights* significativos de grandes volumes de dados textuais. Segundo Kang *et al.* (2020), algumas das aplicações do PLN incluem:

- Análise de Sentimento: O PLN pode ser utilizado para analisar o sentimento expresso em textos relacionados de mercado específicos, como comentários de clientes, análises de mercado e notícias. Isso permite identificar tendências de sentimento positivas ou negativas em relação a determinados produtos, empresas ou eventos, fornecendo *insights* sobre a percepção do mercado.

- Extração de Informações: O PLN pode ser empregado para extrair informações específicas de grandes volumes de texto, como preços de produtos, dados de produção, tendências de mercado e informações sobre concorrentes. Essas informações podem ser utilizadas para monitorar o desempenho do mercado, identificar oportunidades e ameaças e tomar decisões estratégicas informadas.

- Sumarização de Texto: O PLN pode ser utilizado para criar resumos automáticos de textos longos, como relatórios de mercado, análises de tendências e notícias. Isso permite aos analistas obter uma visão rápida e concisa das informações mais relevantes, economizando tempo e facilitando a tomada de decisões.

- Classificação de Texto: O PLN pode ser empregado para classificar textos em categorias ou temas específicos, como tipos de produtos, segmentos de mercado ou tendências emergentes. Isso ajuda a organizar e categorizar grandes volumes de dados textuais, facilitando a análise e a identificação de padrões e *insights* relevantes.

Kang *et al.* (2020)também aborda algumas das principais bibliotecas em *Python* que podem ser utilizadas para PLN, que incluem:

- NLTK (*Natural Language Toolkit*): O NLTK é uma das bibliotecas mais populares para PLN em *Python*. Ele oferece uma ampla gama de ferramentas e recursos para realizar tarefas como "tokenização", *stemming*, "lematização", análise sintática, análise de sentimento e muito mais. O NLTK é uma escolha versátil para uma variedade de aplicações de PLN.

- SpaCy: O SpaCy é outra biblioteca amplamente utilizada para PLN em *Python*. Ele é conhecido por sua velocidade e eficiência, tornando-o ideal para lidar com grandes volumes de texto. O SpaCy oferece funcionalidades como tokenização, análise sintática, reconhecimento de entidades nomeadas e muito mais, sendo uma escolha sólida para análises de textos.

- Gensim: O Gensim é uma biblioteca especializada em modelagem de tópicos e processamento de texto para a extração de informações semânticas de grandes volumes de texto. Ele oferece algoritmos eficientes para a criação e análise de modelos de tópicos, semântica distribuída, similaridade de documentos e outras tarefas relacionadas. O Gensim pode ser útil para identificar padrões e tendências nos dados textuais de divrersos mercados.

- TextBlob: O TextBlob é uma biblioteca simplificada de PLN em *Python*, construída sobre o NLTK e o Pattern. Ele oferece uma interface simples para realizar tarefas comuns de PLN, como tokenização, análise de sentimentos, classificação de texto dentre outros.

Os conjuntos de dados obtidos por *scrapping* tendem a ser extremamente grandes e complexos (*Big data*) e podem ser analisados para revelar padrões, tendências e associações, especialmente relacionados aos comportamentos humanos e suas interações. A abordagem da Inteligência Competitiva (IC) envolve a coleta sistemática, análise e disseminação de informações sobre mercado, concorrentes e o ambiente competitivo, a fim de apoiar a tomada de decisões e o planejamento estratégico dentro de uma organização (RANJAN; FOROPON, 2021).

Kumar, Kar e Ilavarasan (2021) analisaram trabalhos em periódicos de gerenciamento de negócios de renome através de métodos de mineração de texto como Análise de Sentimento, Modelagem de Tópicos e PNL. Foram aplicadas ferramentas de visualização para mineração de texto e associação de tópicos para compreender os temas e relacionamentos dominantes. A análise de Kumar, Kar e Ilavarasan (2021) destacou que a análise das redes sociais, a análise de mercado e a IC como os temas mais dominantes, enquanto outros temas como a gestão de riscos e a detecção de conteúdos falsos também são explorados em menor extensão.

Algoritmos de mineração de texto e machine learning podem ser aplicados em setores como bancos e finanças, *marketing* artístico, varejo e turismo para identificação de segmentos de clientes lucrativos (DEKIMPE, 2020) (PITT; BAL; PLANGGER, 2020). Uma combinação de técnicas de otimização de dados, aprendizado de máquina e árvores causais também pode restringir os clientes-alvo (SIMESTER; TIMOSHENKO; ZOUMPOULIS, 2020).

Nouira, Bouchakwa e Jamoussi (2023) empregaram a análise de sentimento em conjunto com técnicas de *web scraping* para prever o preço do Bitcoin, a aplicação extrai dados de fontes *online* utilizando o *web scraping* para coletar informações relevantes sobre o sentimento do mercado em relação à criptomoeda. O uso de recursos enriquecidos desenvolvidos por meio de processamento de linguagem natural, análise de sentimento e análise de tendências de pesquisa na *web* seguida das análises estatísticas demonstrou forte correlação com o preço do Bitcon.

Segundo Cambria (2016), a emoção é fundamental para a compreensão da preferência humana e do processamento de emoções através da análise de sentimentos, o uso da inteligência artificial pode detectar a polaridade do consumidor. A proliferação cada vez maior de redes sociais exige algoritmos computacionais para dar sentido à *big data* e fornecer aprendizagem profunda sobre os sentimentos polarizados dos consumidores. O conteúdo gerado pelo usuário em sites de redes sociais fornece *insights* profundos do consumidor para melhorar a tomada de decisões (TRIPATHY; AGRAWAL; RATH, 2016).

Ainda, o Matplotlib, biblioteca *Python* mais amplamente utilizada para criação de gráficos, permite representar tendências em dados temporais, distribuições estatísticas ou simplesmente para criar visualizações personalizadas. Segundo Matplotlib Development Team (2022), o ecossistema do Matplotlib é enriquecido com ferramentas complementares, como o Seaborn, que simplifica ainda mais a criação de gráficos estatísticos complexos, e o Pandas, que oferece integração direta para visualização de dados em *DataFrames*, tornando a exploração e a análise de dados mais eficientes e acessíveis para os usuários.

De modo geral, a programação em *Python* oferece um vasto potencial e uma impressionante capacidade devido à sua sintaxe clara e concisa, que facilita a expressão de ideias complexas de forma simples. Sua ampla adoção em uma variedade de campos, desde desenvolvimento *web* e científico até automação e inteligência artificial, destaca sua versatilidade e poder.

Ferramentas digitais desempenham um papel crucial na melhoria do relacionamento com clientes ao oferecerem meios eficazes de comunicação e interação. Plataformas de CRM (*Customer Relationship Management*), por exemplo, permitem às empresas gerenciar dados de clientes de maneira organizada, oferecendo *insights* personalizados e históricos de interações. Além disso, redes sociais e aplicativos de mensagens possibilitam uma comunicação instantânea e direta, facilitando o atendimento ao cliente em tempo real e a resolução ágil de problemas. Scraping e crawling também podem alavancar o relacionamento com clientes ao proporcionar acesso a dados valiosos e *insights* significativos. É possível coletar informações relevantes sobre preferências, comportamentos de compra e *feedback* dos clientes em plataformas *online*. Isso permite às empresas entender melhor as necessidades e interesses dos clientes, adaptando seus produtos e serviços de acordo(SIMESTER; TIMOSHENKO; ZOUMPOULIS, 2020).

Esse monitoramento proativo não só ajuda a gerenciar a reputação da empresa, mas também demonstra um compromisso com a transparência e a prontidão em abordar as preocupações dos clientes, o que pode fortalecer a confiança e a lealdade do cliente (WANG; LAI; LIN, 2023). Assim, o uso estratégico de dados coletados através de *crawling* e *scraping* pode transformar a maneira como uma empresa interage com seus clientes, tornando-se mais responsiva e orientada por dados.

3 Metodologia

3.1 Definição de propósitos e escopo da ferramenta

Para o desenvolvimento desta ferramenta foram considerados aspectos específicos do setor de celulose e papel a qual desejou-se a captura de informações que dissessem respeito ao relacionamento com os clientes e a competitividade das empresas. Buscou-se a utilização de elementos para a identificação de eventos relevantes, tendências de mercado, preferências regionais dos consumidores, análise da concorrência, satisfação dos clientes, percepção de sustentabilidade, coleta de inteligência competitiva e antecipação de tendências futuras. Dessa forma, foram definidas as palavras-chave que contivessem elementos de interesse para a indústria e permitissem a monitorização contínua de fontes online que atendessem aos propósitos da ferramenta.

3.2 Fontes de Dados

Foram determinadas as fontes de dados relevantes para o atendimento dos propósitos definidos no item anterior. Nessa fase, foram consideradas fontes online, principalmente *sites* de notícias, tendo o *Google News* (Google Inc., 2024) como fonte primária de buscas. As fontes utilizadas no trabalho foram verificadas para que pudesse oferecer informações relevantes sobre o setor.

3.3 Aspectos Legais

A *web scrapping* tem sido facilitada através do desenvolvimento de uma variedade de ferramentas e tecnologias. No entanto, as implicações legais e éticas de empregar essas ferramentas para a coleta de dados são frequentemente desconsideradas. A consideração inadequada desses fatores de *scraping* na *Web* pode resultar em grandes disputas e ações éticas (KROTOV; SILVA, 2018).

O cenário legal em torno do tema ainda está se desenvolvendo, e os tribunais estão apenas começando a abordar reivindicações por razões de análise. Definir se rastrear ou raspar os dados para fins de análise levanta problemas legais é uma tarefa altamente específica e dependente de fatos. Não obstante, não existe nenhuma legislatura que trate diretamente do *web scraping* (KROTOV; SILVA, 2018). Segundo Krotov e Johnson (2022), *Web Scraping* tem sido guiado por um conjunto de teorias e leis jurídicas fundamentais relacionadas, como “violação de direitos autorais”, “quebra de contrato”, a Lei de Fraude e Abuso de Computador (CFAA) e “transgressão de bens móveis”.

Do ponto de vista jurídico, uma pergunta que as entidades devem se fazer é se sua ação de *scraping* prejudica o *site* copiado. Se a atividade de raspagem for muito intensa, o que pode interromper os serviços do *site* raspado ou se os dados raspados forem usados de forma a duplicar a atividade ou o serviço desse *site*, mesmo que não existam regulamentações, o *site* teria motivos para entrar com uma ação judicial contra o raspador. Cabe destacar que a maioria dos *sites* contém explicitamente alguma cláusula nos: terms and services condicionando a utilização dos dados (KROTOV; SILVA, 2018).

Do ponto de vista ético, dado que a raspagem de dados já possui muitos casos de uso e fornecedores profissionais, a aplicação desse tipo de framework para fins comerciais tem sido uma prática em franca expansão mundialmente (KROTOV; JOHNSON, 2022).

Cabe destacar ainda que existem práticas recomendadas e técnicas de *web scraping*, segundo Mitchell (2018) que facilitam a carga de tráfego no *site* fonte, como:

- Coletar apenas os dados necessários, determinando o caso de negócios exato e personalizando sua tecnologia de rastreador da web para ele. Isso minimizará o risco de esgotar o *site* copiado com tráfego indesejado.
- Consultar os termos de uso do *site* copiado. Além dos termos de uso comerciais, os *sites* também possuem um arquivo *robots.txt* que inclui informações sobre as permissões do *site* copiado.
- Certificar de mascarar os dados para minimizar a exposição no processamento.
- Certificar de que os dados sejam armazenados de forma segura, não expondo dados copiados ao público.

É inevitável que *scraping* ou *crawling* para fins de análise evolua continuamente, e que os tribunais lutarão com as teorias e fatos legais. Os *Scrapers* e os proprietários de *sites* devem estar cientes das decisões precedentes e respectivas jurisprudências e cuidadosos desenvolvimentos corrente e em potencial (SNELL; MENALDO, 2016).

3.4 Desenvolvimento do algoritmo

O desenvolvimento do algoritmo em *Python* para pesquisa de notícias baseada em palavras-chave exigiu a criação de funções que permitam aos usuários inserir consultas de forma simples e clara. A definição de parâmetros de busca, como palavras-chave e intervalos de datas para varredura de dados, por meio de variáveis ou argumentos de função em uma estrutura de código bem organizada e comentada direcionou para a compreensão e a interação fluida com o algoritmo.

Além da entrada de dados, foi necessário implementar um mecanismo de crawling para acessar e extrair notícias relevantes de fontes como o *Google News*. Isso pode ser alcançado utilizando biblioteca *Python* especializada em *web scraping* Requests-html. Não obstante, o algoritmo descrito, operou dentro dos limites dos termos de uso dos *sites* de onde as notícias foram extraídas, evitando solicitações excessivas que possam levar a bloqueios, sobrecarga ou penalidades.

Ademais, verificou-se a necessidade de criar uma função de exportação que converta os resultados da pesquisa em formato JSON. Dessa forma, os usuários seriam capazes de salvar os dados obtidos para uso posterior ou análise em outras ferramentas ou sistemas.

Segundo Booch, Rumbaugh e Jacobson (1999), o diagrama de classes representa a estrutura e as relações entre as classes em um sistema orientado a objetos. Ele fornece uma visualização visual das classes, seus atributos e métodos, bem como as associações entre elas. O diagrama de classes permite aos desenvolvedores, durante a modelagem de sistemas de *software*, entenderem a arquitetura do sistema, organizar e modularizar o código de forma eficiente e comunicar as ideias e *design* do sistema para outras partes interessadas.

Está representado pela Figura 1, o diagrama de classes do sistema que realiza *web crawling* e *web scraping* a partir de consultas inseridas no *Google News*, seguido pela análise de polaridade de sentimentos, análise de multi-sentimentos e suas respectivas representações gráficas, além da construção da nuvem de palavras, dividido em várias classes inter-relacionadas para modularizar as funcionalidades do sistema.

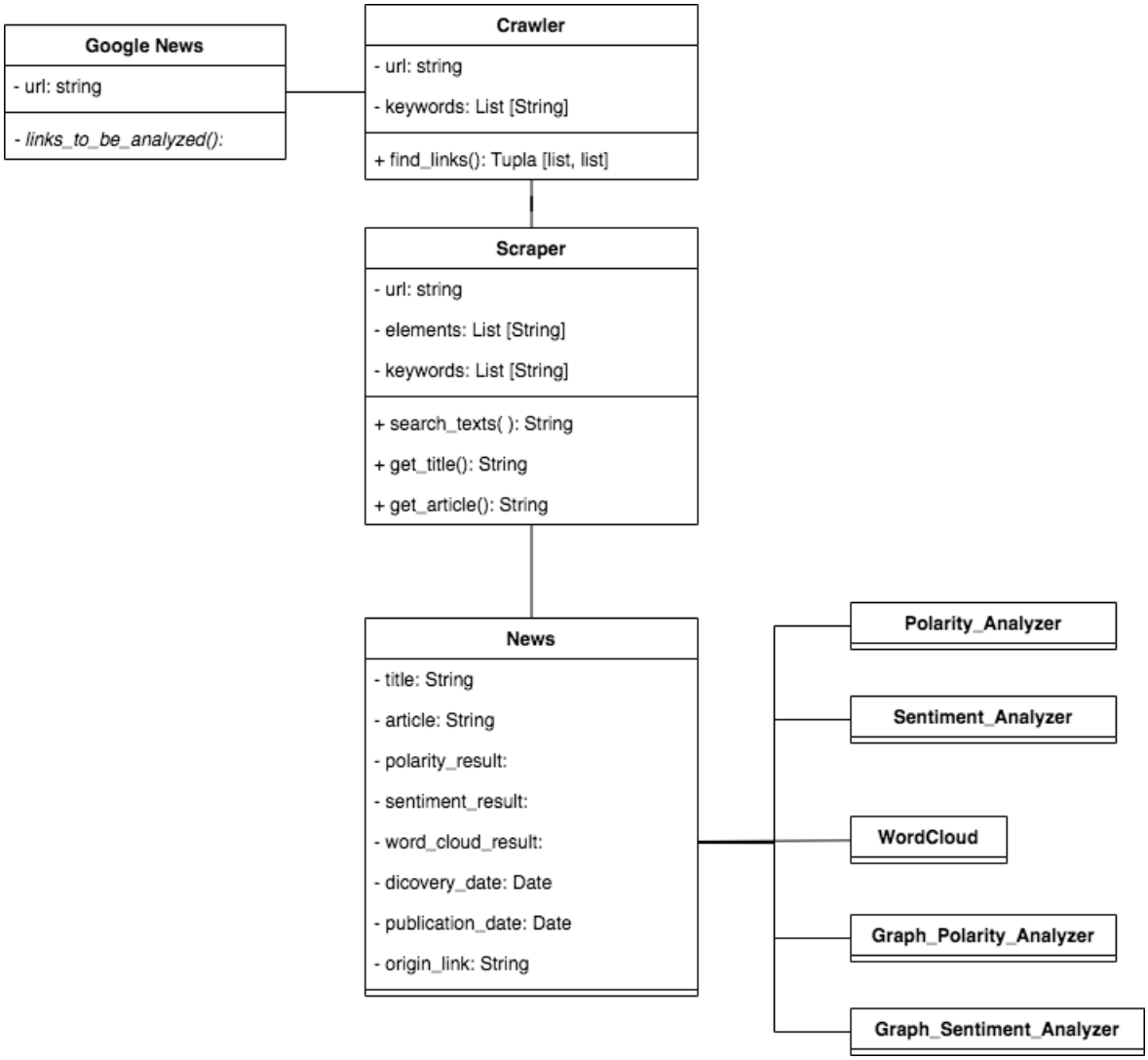
A classe central é o "*Crawler*", responsável por iniciar o processo de busca no *Google News* a partir da inserção da(s) palavra(s)-chave e coordenar as operações subsequentes. Ela interage com a classe de "*Scraper*"para extrair os dados relevantes das páginas web retornadas pela pesquisa. Nesse ponto, a palavra chave é comparada a um grupo de palavras previamente selecionadas para comparação da existência dela nos textos das notícias.

A classe "*News*"representa cada notícia coletada, contendo informações como título, conteúdo, url, data, entre outros. Essa classe tem a finalidade de armazenar os dados extraídos durante o processo de *web scraping* e serve como entrada nas classes subsequentes.

Após a extração dos dados, as informações são passadas para as classes de "*Polarity Analyzer*"e "*Sentiment Analyzer* ", responsáveis por realizar a análise de polaridade de sentimentos e multi-sentimento respectivamente.

Dentro da classe "*Polarity Analyzer*", o método para análise de polaridade de sentimentos determinaria se uma notícia é positiva, neutra ou negativa, com a combinação final dos “polos” do texto da notícia. A "*Sentiment Analyzer*"propõe uma análise de multi-sentimentos, que categorizaria as notícias em diferentes emoções como alegria, surpresa, tristeza, nojo e raiva.

Figura 1 – Diagrama de classes para representação do sistema proposto



Fonte: Elaborado pelo autor.

Paralelamente, a "*News*" é utilizada para extrair as palavras-chave mais relevantes de cada notícia. Esse processo é conduzido de métodos de processamento de texto, como tokenização, remoção de stopwords e contagem de frequência de palavras. Conforme Bird, Klein e Loper (2009), após a coleta de dados, é essencial a realização do pré-processamento para extrair informações mais relevantes. As palavras-chave são transferidas para a classe de "*Word Cloud*"para criar a nuvem de palavras correspondente.

Quanto à plotagem dos resultados das análises de sentimentos, foram aplicados métodos na classe "*Polarity Analyzer*"e "*Sentiment Analyzer*" para gerar gráficos individualizados para a polaridade de sentimentos e para cada categoria de multi-sentimento. Foram utilizadas a biblioteca de visualização de dados, Matplotlib (Matplotlib Development Team, 2022), para criar gráficos do tipo *3D Scatterplot* dos dados obitdos em "*Polarity Analyzer*" e para criação de gráficos de barras horizontais de distribuição discreta dos dados obitidos em "*Sentiment Analyzer*" como forma de representação da distribuição dos diferentes tipos de sentimentos encontrados nas notícias.

Essas classes em trabalho conjunto forneceram, portanto, uma solução para a busca, coleta e análise de notícias do *Google News*, que tende a condicionar uma compreensão do sentimento associado às informações encontradas.

3.5 Estruturação da base de dados

JSON, ou *Java Script Object Notation*, é um formato leve de troca de dados que é fácil de ler e escrever para humanos e fácil de analisar e gerar para máquinas. Ele é baseado em um subconjunto da linguagem de programação JavaScript, mas é independente de linguagem, tornando-se uma escolha popular para transferir dados entre um servidor e um cliente da *web*. Em sua forma mais básica, o JSON consiste em pares de chave-valor, onde as chaves são strings e os valores podem ser qualquer tipo de dados válido, incluindo strings, números, objetos, *arrays*, booleanos e valores nulos. Este formato é comumente usado em APIs da *web* para enviar dados estruturados entre diferentes sistemas e aplicativos (PENG; CAO; XU, 2011).

Sendo assim, exportar dados estruturados de notícias de uma ferramenta criada para *scraping* do *Google News* em formato JSON facilitou o processamento e a análise dos dados por parte de outros sistemas que podem ser facilmente consumidos e utilizados em uma variedade de contextos, como análises de dados e visualizações. Outro ponto que ponderou a favor da utilização deste formato, foi a capacidade de manter a estrutura e a integridade dos dados originais. O JSON permitiu representar dados de maneira hierárquica e organizada, preservando a relação entre os diferentes elementos das notícias, como título, url fonte, data de publicação, conteúdo e assim por diante. Dessa forma, maior compreensão e a facilidade na interpretação dos dados exportados foi possível, garantindo qualidade e consistência dos dados.

3.6 Documentação e registro

Documentar o código, os processos e as decisões de design para facilitar futuras atualizações e manutenção foi parte importante para a construção da ferramenta. Para o presente trabalho, foi utilizada a plataforma de hospedagem de código-fonte baseada na *web* GitHub (TORVALDS, n.d.) para gerenciamento do projeto e controle de versão.

Outras partes relevantes do processo de documentação se referem a integração de feedback dos usuários e a realização de ajustes conforme necessário. Estes seguiram princípios ágeis descritos por Beck *et al.* (2001). Testes extensivos foram realizados para garantir a precisão, eficiência e usabilidade da ferramenta.

4 Resultados

4.1 Recuperação de Informações

A solução proposta para extrair e estruturar as informações das notícias no formato JSON alcançou resultados esperados para a recuperação de informações. Através de técnicas de coleta de dados, a solução foi capaz de acessar uma ampla variedade de fontes de notícias online, proporcionando uma base de dados abrangente e atualizada. A estruturação das informações em formato JSON permitiu uma fácil manipulação e análise dos dados, facilitando a extração de insights relevantes para diversas aplicações, desde análises de tendências até a geração de relatórios detalhados sobre temas específicos (Figura 2).

Figura 2 – Exemplo de dados de notícia extraído no formato Json.


```
  "title": "Fibra de celulose é alternativa para uso na construção civil",
  "origin_link":
"https://news.google.com/articles/CBMibWh0dHBzOi8vd3d3Lmdvdi5ici9jYXBlcy9wdC1
ici9hc3N1bnRvcy9ub3RpY2lhcy9maWJyYS1kZS1jZWx1bG9zZS1lLWFsdGVybmF0aXZhLXBhcmEt
dXNvLW5hLWNvbnN0cnVjYW8tY2l2aWzSAQA?hl=pt-BR&gl=BR&ceid=BR%3Apt-419",
  "publication_date": "2023-12-13T08:00:00Z",
  "discovery_date": "2024-04-24T12:34:40Z",
  "article": "...A Associação Brasileira Técnica de Celulose e Papel atua como
uma ponte de conexão, promovendo a troca de conhecimentos, experiências e
oportunidades entre esses três pilares. A ABTCP oferece diversos recursos e
programas voltados para estudantes, como eventos técnicos, congressos,
workshops e cursos de capacitação. Essas iniciativas proporcionam um ambiente
propício para que os estudantes possam estar em contato direto com profissionais
da indústria, conhecer as últimas tendências e avanços tecnológicos, além de
expandir sua rede de contatos.",
```


Fonte: Elaborado pelo autor.

Além disso, a solução demonstrou uma eficiência na recuperação precisa de informações, minimizando a ocorrência de dados inconsistentes ou irrelevantes. A aplicação de algoritmos de filtragem e validação, permitiu a obtenção de dados com a qualidade e a confiabilidade.

4.2 Análises de Sentimentos

Para a análise dos dados, a aplicação de algoritmos de *machine learning* pode ser considerada (RASCHKA; MIRJALILI, 2017). A incorporação de recursos de processamento de linguagem natural (PLN), como análise de sentimentos e extração de tópicos, baseia-se em técnicas modernas descritas por Jurafsky e Martin (2019) em "*Speech and Language Processing*". Estas técnicas podem aprimorar significativamente a análise e interpretação dos dados coletados.

A interpretação das informações extraídas do *Google News* do setor de celulose e papel utilizou a análise de sentimento da biblioteca *Natural Language Toolkit* (NLTK) (Bird, Steven and Loper, Edward and Klein, Ewan, 2009). Primeiramente, a análise de sentimento permitiu identificar a polaridade das opiniões expressas nas notícias, classificando-as como positivas, negativas ou neutras. Ao analisar um conjunto de notícias sobre o setor de celulose e papel, a análise de sentimento pode revelar, para um dado instante, se a cobertura da mídia era predominantemente positiva, negativa ou equilibrada em relação a questões como desempenho do mercado, inovações tecnológicas, questões ambientais e regulatórias, entre outros.

O *Sentiment Analyzer* da NLTK oferece suporte a diferentes idiomas. Dessa forma, foi possível a aplicações em contextos no idioma Português para a solução proposta. Além disso, foi possível realizar uma série de tarefas relacionadas à análise de sentimentos em textos além da caracterização da polaridade do sentimento expresso em um determinado texto.

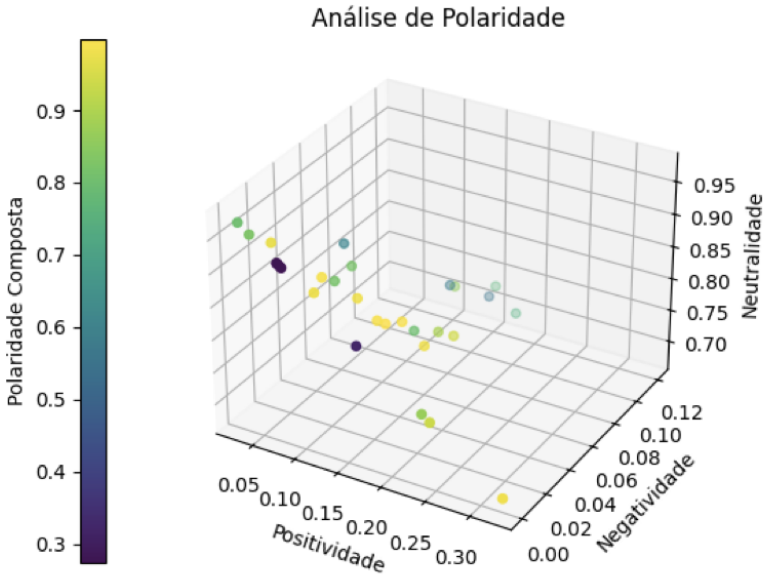
Foi também possível treinar o algoritmo (*Sentiment Analyzer*) com conjuntos de textos que representam diferentes emoções (alegria, surpresa, raiva, aversão, medo e tristeza) a partir dos quais foram criados modelos de análise de sentimentos mais precisos e eficazes para diferentes assuntos. Para o conjunto de textos que representam as diferentes emoções foram utilizados o ChatGPT (OpenAI, 2022), ferramenta baseada em inteligência artificial, e coletados com as emoções correspondentes, para que os dados de treinamento fossem rotulados. O algoritmo submeteu os dados a um pré-processamento dos Dados para remover ruídos e padronizar o formato.

Em seguida, foram extraídas as características dos textos que poderiam ser usadas para treinar o classificador de sentimentos, tais como contagem de palavras-chave relacionadas a cada emoção, padrões de pontuação, etc. Foi utilizado o algoritmo de classificação, *Naive Bayes* da biblioteca NLTK, para criar um modelo que fosse capaz de prever as emoções com base nos textos.

Após treinar o classificador, ajustes e otimizações do modelo com base nos resultados de avaliações foram realizados. A avaliação envolveu a experimentação com diferentes conjuntos de características, algoritmos de classificação e parâmetros de ajuste fino para melhorar o desempenho do modelo.

São represantados pelas Figuras 3, 4 e 5 os resultados obtidos ao rodar a solução para a palavra-chave “Celulose”, realizada em 23/04/2024 para notícias a partir do ano de 2024, fundamentada no progresso e na complexidade do desenvolvimento da ferramenta.

Figura 3 – Distribuição das polaridades obtidas pela Análise de Sentimento (*Polarity Analyzer*) referentes a 29 primeiras notícias durante busca em 23/04/2024.



Fonte: Elaborado pelo autor.

O classificador de sentimento (Figura 3) representa um conjunto de 29 primeiras notícias relacionadas ao termo de pesquisa "celulose". Cada ponto na figura representa uma notícia. A resposta da solução revelou que a maioria das notícias estava dentro do espectro de sentimentos positivos. Isso sugere uma tendência predominante de reportagens favoráveis sobre a indústria da celulose, que podem abordar avanços tecnológicos, novas oportunidades de negócios e investimentos ou impactos positivos na sociedade de modo geral como geração de emprego e renda.

Cabe destacar a importância da análise de sentimento na compreensão da percepção pública e na monitorização da reputação de determinados setores industriais. Além disso, o resultado sugere um clima favorável e otimista em relação à celulose, pelo menos momentaneamente. Esse tipo de resultado (ou conjunto de resultados) pode eventualmente ser escalável para influenciar decisões de investimento, estratégias de comunicação e políticas governamentais relacionadas ao setor.

Figura 4 – Resultados da Análise de Multi sentimentos (*Sentiment Analyzer*) extraídas das 29 primeiras notícias da solução proposta. Palavra chave: Celulose.

Fonte: Elaborado pelo autor.

Na classificação por polaridade, os textos são geralmente rotulados como positivos, negativos ou neutros, com base na polaridade geral do sentimento expresso. Essa abordagem tende a simplificar a interpretação, fornecendo uma visão geral do sentimento predominante em um texto. No entanto, ela pode perder nuances e complexidades, especialmente em textos que contenham opiniões mistas ou sentimentos sutis. Por outro lado, na classificação multi-sentimentos, os textos são classificados em várias categorias que capturam uma gama mais ampla de sentimentos e nuances. Isso pode incluir categorias como "alegria", "medo", "aversão", entre outras. Essa abordagem permite uma interpretação mais granular e detalhada dos sentimentos expressos nos textos, capturando variações na intensidade e na natureza dos sentimentos.

Enquanto a classificação por polaridade oferece uma visão mais simplificada e direta dos sentimentos, a classificação multimodal fornece uma representação mais rica e detalhada, o que pode levar a interpretações diferentes, especialmente em textos que contenham uma variedade de opiniões e sentimentos. A escolha entre essas abordagens depende das necessidades específicas da análise de sentimentos e das nuances do contexto em questão.

Em todo caso, cabe destacar a coerência entre o espectro positivo das notícias extraídas e avaliadas na análise de polaridade e a demonstração de sentimentos mais “positivos” na análise de multi sentimentos

4.3 *Word Cloud*

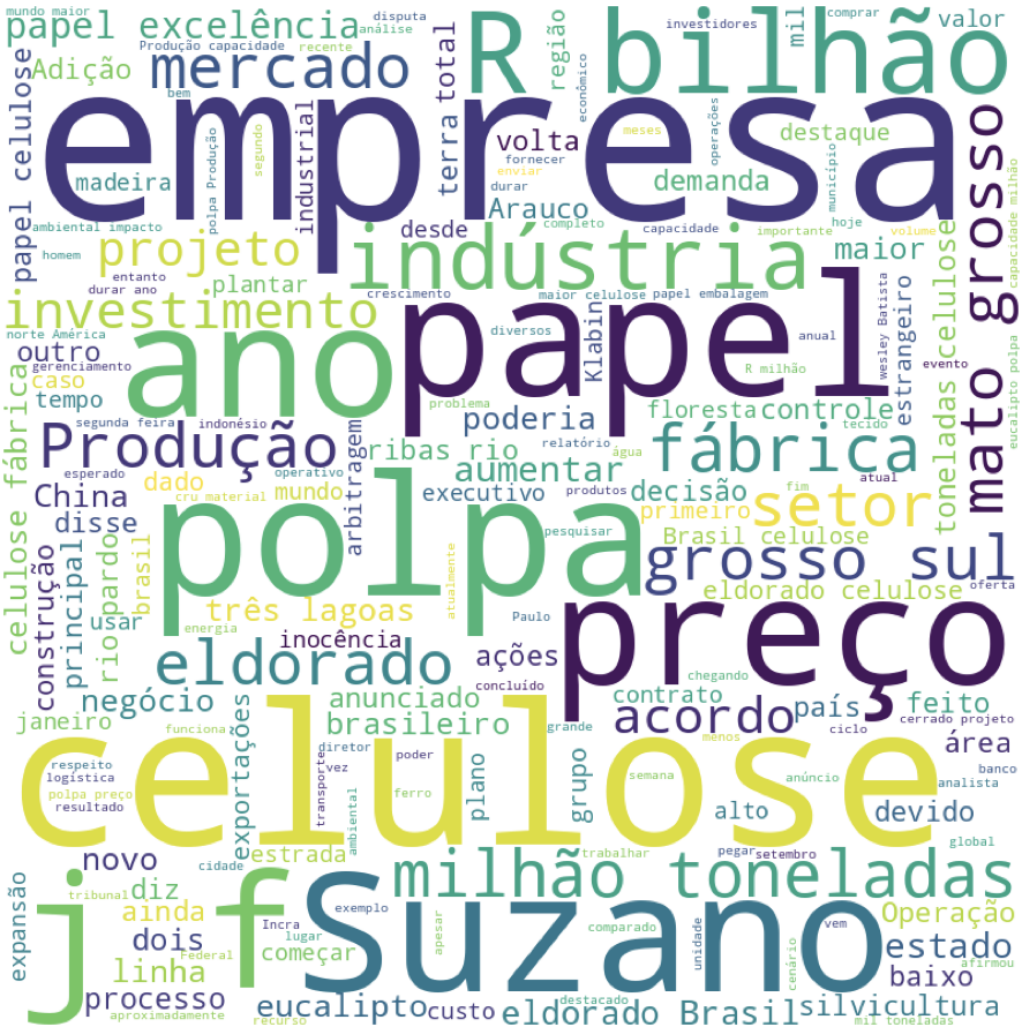
As nuvens de palavras são populares para análise de *sites* e textos. Numa análise de texto típica, palavras de interesse (por exemplo, de um documento, títulos de periódicos ou tags de um *site*) são colocadas em formato retangular. O tamanho da fonte e a cor das palavras são colocadas na nuvem de palavras para representar frequência e utilidade, respectivamente. Outras opções, como estilo de fonte e *layout*, estão disponíveis para aprimorar o apelo visual das nuvens de palavras. Em uma nuvem de palavras típica, as tags de um *site* (ou palavras de um documento) são compactadas em uma região retangular na qual o tamanho da fonte indica a popularidade da tag (ou frequência das palavras) e a cor da fonte indica outras informações úteis (CUI *et al.*, 2010). Quanto mais proeminente (tamanho de texto maior) a palavra estiver na nuvem de palavras, mais frequentemente ela aparecerá no texto fornecido.

O objetivo das nuvens de palavras é resumir termos importantes em uma apresentação visual que ajuda a sintetizar as “grandes ideias” presentes, sejam elas relacionadas ao conteúdo importante do curso (HAMM, 2011) ou a questões relacionadas à cirurgia (MCGEE; MCGEE, 2011). Ao contrário de simplesmente listar os termos mais importantes com uma contagem de frequência em outra forma, como uma tabela, as nuvens de palavras nestas e em outras situações de Nuvem de Palavras oferecem um gráfico que oferece “*clusters* semanticamente significativos com *layouts* visualmente atraentes” (CUI *et al.*, 2010).

Extrair e interpretar informações de uma nuvem de palavras de uma notícia do setor de celulose e papel envolve um processo de análise cuidadosa dos termos mais frequentes no texto. Por exemplo, se analisarmos uma nuvem de palavras relacionada a uma notícia sobre a indústria de celulose, termos como "produção", "aumentar", "preço", "demanda", "mercado"e "China"podem surgir com destaque (Figura 5). Essas palavras indicam áreas-chave de interesse e foco nas notícias, como a produção e exportação de celulose, os preços no mercado internacional, o aumento anunciado da produção de celulose no estado do Mato Grosso e a demanda por produtos de celulose, especialmente na China.

Dentre as notícias, houve destaque aos avanços e investimentos significativos no setor da celulose. A manchete envolve a Eldorado Brasil, cujo anúncio da construção de uma nova linha de produção, com investimento de R$ 25 bilhões, posiciona Três Lagoas no Mato Grosso do Sul como um polo mundial da celulose. A iniciativa contribuirá para a consolidação da região como um importante centro de produção e exportação de celulose, impulsionando a economia local e nacional (NEWS, 2024).

Figura 5 – Exemplo da nuvem de palavras criada a partir da busca da palavra celulose na solução proposta.



Fonte: Elaborado pelo autor.

Outra notícia em destaque relata que a empresa Suzano enviou à China a maior carga de celulose já transportada em um navio. A carga foi embarcada em Barcarena, no Pará, e transportada pelo navio Cap San Artemissio, da Hamburg Süd. A operação representa um marco para a indústria, destacando a capacidade logística da empresa em atender à demanda global por celulose. A carga totalizou cerca de 77 mil toneladas e é destinada ao mercado chinês, demonstrando a importância desse país como destino para os produtos (ISTOÉ DINHEIRO, 2024).

O Ministério da Educação, em uma notícia datada de 21 de abril de 2024, discute o potencial da fibra de celulose como alternativa para uso na construção civil. O artigo destaca os benefícios ambientais e econômicos dessa alternativa, bem como seu papel na redução da pegada de carbono da indústria da construção. A fibra de celulose é apresentada como um material renovável e sustentável, capaz de substituir produtos tradicionais, como o concreto, em várias aplicações na construção. O texto também destaca iniciativas de pesquisa e desenvolvimento que buscam explorar ainda mais o potencial da fibra de celulose nesse contexto (Ministério da Educação, 2024).

Essas notícias destacam avanços significativos na indústria de celulose, especialmente no Brasil. Sendo assim, ao interpretar tanto a nuvem de palavras quanto às análises de sentimentos e polaridade, deve-se levar em conta fatores como tendências de mercado, regulamentações governamentais, desafios ambientais e avanços tecnológicos que podem influenciar a indústria como um todo. Ao fazer isso, podemos extrair insights valiosos sobre as principais preocupações, tendências e oportunidades que estão moldando o setor de celulose e papel e informar decisões estratégicas para empresas e profissionais do ramo, uma vez que é crucial a verificação do contexto global do setor dentro do referido instante da criação da nuvem.

5 Considerações Finais

O desenvolvimento da ferramenta para busca e indexação de assuntos relacionados ao setor de celulose e papel conduziu a um avanço na capacidade de obter insights valiosos para auxiliar no processo de tomada de decisão nessa indústria crucial. Por meio da aplicação de técnicas de crawling, scraping e processamento de linguagem natural, a ferramenta permite uma análise abrangente e em tempo real dos produtos, tópicos e principais mercados associados ao setor.

Ao empregar técnicas avançadas de crawling, a ferramenta é capaz de coletar dados de fontes diversas, tendo o *Google News* como principal fonte para extração de dados relevantes para o setor de celulose e papel. Dessa forma, garantiu-se uma ampla cobertura de informações, fornecendo uma base sólida para análises posteriores.

Além disso, ao utilizar técnicas de scraping e processamento de linguagem natural, a ferramenta foi capaz de extrair *insights* significativos dos dados coletados, identificando tendências, padrões e correlações que podem ser úteis para os tomadores de decisão no setor para uma compreensão mais profunda e informada do ambiente de negócios, facilitando a identificação de oportunidades, caracterizando seu potencial em detectar ameaças e áreas de foco estratégico.

Em última análise, a ferramenta pode certamente oferecer uma vantagem competitiva ao fornecer informações acionáveis e relevantes para orientar o desenvolvimento de estratégias eficazes e a tomada de decisões informadas na indústria de celulose e papel.

Referências

ALDEEN, Y. A. A. S.; SALLEH, M.; RAZZAQUE, M. A. A comprehensive review on privacy preserving data mining. **SpringerPlus**, v. 4, p. 694, 2015.

ARJUNAN, T. Building business intelligence data extractor using nlp and python. **International Journal of Innovative Science and Research Technology**, v. 7, n. 9, 2022.

AWAN *et al.* Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance. **Technological Forecasting and Social Change**, v. 168, p. Article 120766, 2021.

BECK, K.; BEEDLE, M.; BENNEKUM, A. V.; COCKBURN, A.; CUNNINGHAM, W.; FOWLER, M.; KERN, J. Manifesto for agile software development. 2001.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. [S.l.]: O'Reilly Media, 2009.

Bird, Steven and Loper, Edward and Klein, Ewan. **Natural Language Toolkit (NLTK)**. [S.l.], 2009. Disponível em: <https://www.nltk.org/>.

BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. **Unified Modelling Language User Guide, The (2nd Edition)**. [S.l.]: Addison-Wesley, 1999.

BROUCKE, S. V.; BAESENS, B. From web scraping to web crawling. In: **Practical Web Scraping for Data Science**. [S.l.: s.n.], 2018. p. 155–172.

CAMBRIA, E. Affective computing and sentiment analysis. **IEEE Intelligent Systems**, v. 31, n. 2, p. 102–107, 2016.

CHEN, H. *et al.* Big data: A survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2012.

CUI, W. *et al.* Context-preserving, dynamic word cloud visualization. **IEEE Computer Graphics & Applications**, v. 30, n. 6, p. 42–53, 2010.

DAVENPORT, T. H.; HARRIS, J. **Competing on Analytics: The New Science of Winning**. [S.l.]: Harvard Business Press, 2007.

DEKIMPE, M. Retailing and retailing research in the age of big data analytics. **International Journal of Research in Marketing**, v. 37, p. 3–14, 2020.

Google Inc. **Google News**. 2024. Disponível em: <https://news.google.com/>.

HAMM, S. E. Using word clouds for reflection and discussion in an online class. **Teaching Theology & Religion**, v. 14, n. 2, p. 156, 2011.

Indústria Brasileira de Árvores. **Relatório Anual 2022**. 2022. Recuperado em 27/03/2024. Disponível em: <https://iba.org/datafiles/publicacoes/relatorios/relatorio-anual-iba2022-compactado.pdf>.

ISTOÉ DINHEIRO. **Suzano envia à China maior carga de celulose já transportada em navio no mundo**. 2024. Disponível em: <https://istoedinheiro.com.br/suzano-envia-a-china-maior-carga-de-celulose-ja-transportada-em-navio-no-mundo/>.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3rd. ed. [S.l.]: Pearson, 2019.

KANG, Y. *et al.* Natural language processing (nlp) in management research: A literature review. **Journal of Management Analytics**, v. 7, n. 2, p. 139–172, 2020.

KROTOV, V.; JOHNSON, L. Big web data: Challenges related to data, technology, legality, and ethics. **Business Horizons**, 2022.

KROTOV, V.; SILVA, L. Legality and ethics of web scraping. In: **Twenty-fourth Americas Conference on Information Systems**. [S.l.: s.n.], 2018.

KUMAR, S.; KAR, A.; ILAVARASAN, P. Applications of text mining in services management: A systematic literature review. **International Journal of Information Management Data Insights**, v. 1, n. 1, 2021.

Matplotlib Development Team. **Matplotlib: Python plotting**. [S.l.], 2022. Disponível em: <https://matplotlib.org/>.

MCAFEE, A.; BRYNJOLFSSON, E. Big data: The management revolution. **Harvard Business Review**, v. 90, n. 10, p. 60–68, 2012.

MCGEE, R. G.; MCGEE, L. M. A picture is worth a thousand words. **American Journal of Transplantation**, v. 11, n. 4, p. 871–872, 2011.

Ministério da Educação. **Fibra de celulose é alternativa para uso na construção civil**. 2024. <https://www.gov.br/capes/pt-br/assuntos/noticias/fibra-de-celulose-e-alternativa-para-uso-na-construcao-civil>.

MITCHELL, R. **Web Scraping with Python: A Comprehensive Guide**. [S.l.]: O'Reilly Media, 2018.

NEWS, C. G. **Eldorado prevê investir R$ 2,5 bilhões em 2ª fábrica de celulose em Três Lagoas**. 2024. Disponível em: <https://www.campograndenews.com.br/economia/eldorado-preve-investir-r-25-bilhoes-em-2a-fabrica-de-celulose-em-tres-lagoas>.

NOUIRA, A. Y.; BOUCHAKWA, M.; JAMOUSSI, Y. Bitcoin price prediction considering sentiment analysis on twitter and google news. In: **Proceedings of the 27th International Database Engineered Applications Symposium**. [S.l.: s.n.], 2023. p. 71–78.

OpenAI. **ChatGPT**. 2022. Disponível em: <https://openai.com/chatgpt>.

PENG, D.; CAO, L.; XU, W. Using json for data exchanging in web service applications. **Journal of Computational Information Systems**, v. 7, n. 16, p. 5883–5890, 2011.

PITT, S.; BAL, S.; PLANGGER, K. New approaches to psychographic consumer segmentation: Exploring fine art collectors using artificial intelligence, automated text analysis and correspondence analysis. **European Journal of Marketing**, 2020.

RANJAN, J.; FOROPON, C. Big data analytics in building the competitive intelligence of organizations. **International Journal of Information Management**, v. 56, 2021.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**. [S.l.]: Packt Publishing, 2017.

REITZ, K. **request-html: Pythonic HTML Parsing for Humans™**. 2019. <https://github.com/psf/requests-html>.

RICHARDSON, L. **Beautiful Soup Documentation**. 2021. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.

Scrapy Developers. **Scrapy**. acessado em 2024. <https://scrapy.org/>.

SIMESTER, D.; TIMOSHENKO, A.; ZOUMPOULIS, S. I. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. **Management Science**, v. 66, n. 6, p. 2495–2522, 2020.

SNELL, J.; MENALDO, N. Web scraping in an era of big data 2.0. **Bloomberg Law News**, 2016.

THOMAS, D. M.; MATHUR, S. Data analysis by web scraping using python. In: **2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)**. [S.l.: s.n.], 2019.

TORVALDS, L. **GitHub**. n.d. Disponível em: <https://github.com/>.

TRIPATHY, A.; AGRAWAL, A.; RATH, S. K. Classification of sentiment reviews using n-gram machine learning approach expert systems with applications. **Expert Systems with Applications**, 2016.

WANG, J.; LAI, J.-Y.; LIN, Y.-H. Social media analytics for mining customer complaints to explore product opportunities. **Computers & Industrial Engineering**, v. 178, 2023.