UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Aplicação de Aprendizado de Máquina na Otimização de Triagem de Orçamentos para Propostas de Vendas de Válvulas Industriais

**João Carlos da Rosa e Silva Vitorino**

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

**João Carlos da Rosa e Silva Vitorino**

# Aplicação de Aprendizado de Máquina na Otimização de Triagem de Orçamentos para Propostas de Vendas de Válvulas Industriais

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

**Versão original**

**São Carlos**
**2023**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

V845a

VITORINO, JOAO CARLOS DA ROSA E SILVA
Aplicação de Aprendizado de Máquina na Otimização de Triagem de Orçamentos para Propostas de Vendas de Válvulas Industriais / JOAO CARLOS DA ROSA E SILVA VITORINO; orientador Dr. Ricardo Rodrigues Ciferri. -- São Carlos, 2023.
56 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Aprendizado de Máquina. 2. Machine Learning. 3. Otimização da triagem de propostas para vendas técnicas. 4. Válvulas Industriais. I. Ciferri, Dr. Ricardo Rodrigues, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**João Carlos da Rosa e Silva Vitorino**

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.
Concentration area: Artificial Intelligence
Advisor: Prof. Dr. Ricardo Rodrigues Ciferri

**Original version**

**São Carlos**
**2023**

*Dedico este trabalho àqueles que escolhem o caminho do esforço e da perseverança em busca do conhecimento e da realização pessoal. Àqueles que, independentemente da idade ou das circunstâncias, se dedicam incansavelmente aos seus objetivos, rejeitando a inércia e a complacência. Este é um tributo aos que acreditam no poder transformador da educação e na meritocracia como reconhecimento justo do trabalho duro e da determinação. Àqueles que, mesmo diante dos obstáculos, mantêm a fé e a esperança, provando que o sucesso é fruto da resiliência e da dedicação incansável.*

## AGRADECIMENTOS

Primeiramente, expresso minha profunda gratidão a Deus, cuja presença e guia foram fundamentais em cada passo desta jornada. Sua força e luz me sustentaram nos momentos de desafio e celebração.

Dirijo um agradecimento especial à minha família, em especial à minha esposa Bianca, e aos meus filhos, João, Luís e Maria. O apoio, incentivo e compreensão de vocês foram inestimáveis. Agradeço por estarem sempre ao meu lado, oferecendo-me amor e encorajamento, sendo a retaguarda sólida em todos os momentos que precisei.

Dedico um tributo carinhoso à memória de minha querida mãe, Therezinha, cujo legado e crença na transformação pela educação continuam a me inspirar. Ela sempre acreditou que a educação tem o poder de mudar vidas e seu exemplo e ensinamentos permanecem como uma luz guia em minha jornada.

"Aprender é a única coisa que a mente nunca se cansa, nunca tem medo e nunca se arrepende."

Leonardo Da Vinci

"Inteligência é a habilidade de se adaptar às mudanças."

Stephen Hawking

"O verdadeiro conhecimento existe em conhecer que você nada sabe."

Sócrates

## RESUMO

VITORINO, J.C.R.S**. Aplicação de Aprendizado de Máquina na Otimização de Triagem de Orçamentos para Propostas de Vendas de Válvulas Industriais**. 2023. 71p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

No cenário comercial contemporâneo, o processo de elaboração de propostas e cotações é de suma importância, atuando como um mecanismo essencial na construção e fortalecimento de relações duradouras com os clientes. Uma proposta meticulosamente elaborada não apenas eleva a reputação e credibilidade da empresa no mercado, mas também a posiciona de forma destacada em um ambiente repleto de concorrência. Diante da crescente demanda por eficiência, a automação desse processo torna-se não apenas desejável, mas essencial, com o intuito de otimizar as vendas e enriquecer a experiência oferecida ao cliente. No âmbito industrial, é uma prática recorrente que departamentos de compras, em sinergia com equipes de engenharia, encaminhem listas detalhadas de equipamentos aos fabricantes. No entanto, a frequente ausência de especificações claras e precisas nessas listas impõe desafios adicionais aos fabricantes, que se veem na tarefa de filtrar e selecionar equipamentos que se alinham ao seu portfólio. Esta prática, muitas vezes adotada devido às limitações de tempo ou recursos, amplia a responsabilidade dos fabricantes em garantir a total conformidade das especificações. Especificamente no segmento de válvulas industriais, a ampla variedade e complexidade dos produtos, aliadas a orçamentos que muitas vezes incluem itens fora do escopo principal e apresentam uma natureza semiestruturada, tornam o processo de triagem uma tarefa complexa, com implicações diretas na eficiência e nos custos operacionais. Diante deste panorama, este estudo visa aplicar técnicas avançadas de aprendizado de máquina, com ênfase em Processamento de Linguagem Natural (NLP), para identificar, categorizar e classificar de forma automática palavras-chave pertinentes a válvulas. A meta é aprimorar e agilizar o processo de triagem, reduzindo o tempo de análise de propostas e garantindo uma seleção mais precisa, especialmente quando confrontados com orçamentos de natureza variada e, por vezes, semiestruturada.



**Palavras-chave**: 1. Aprendizado de Máquina. 2. Machine Learning. 3. Otimização da triagem de propostas para vendas técnicas. 4. Válvulas Industriais.

**ABSTRACT**

VITORINO, J.C.R.S. **Application of Machine Learning in the Optimization of Budget Screening for Industrial Valve Sales Proposals.** 2023. 71p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

In the contemporary commercial landscape, the process of drafting proposals and quotations stands as a fundamental importance, serving as a pivotal mechanism in building and reinforcing lasting relationships with clients. A meticulously crafted proposal not only boosts a company's market reputation and credibility but also sets it apart in a fiercely competitive environment. Faced with the growing demand for efficiency, automating this process has transitioned from being merely desirable to essential, aiming to enhance sales and enrich the client experience. Within the industrial context, it's a recurring practice for purchasing departments, in collaboration with engineering teams, to send detailed equipment lists to manufacturers. However, the often-observed lack of clear and precise specifications in these lists presents additional challenges to manufacturers, who find themselves tasked with filtering and selecting equipment that aligns with their portfolio. This approach, frequently adopted due to time or resource constraints, heightens the manufacturer's responsibility to ensure full specification compliance. Particularly in the industrial valve sector, the broad variety and intricacy of products, combined with budgets that often encompass items beyond the primary scope and exhibit a semi-structured nature, render the screening process a multifaceted task, with direct implications on operational efficiency and costs. In light of this scenario, this study aims to leverage advanced machine learning techniques, emphasizing Natural Language Processing (NLP), to automatically identify, categorize, and classify valve-related keywords. The objective is to refine and expedite the screening process, curtailing proposal analysis time and ensuring a more accurate selection, especially when faced with budgets of diverse and occasionally semi-structured character.

**Keywords:** 1. Machine Learning. 2. Optimization of proposal screening for technical sales. 3. Industrial Valves.

## LISTA DE BREVIATURAS E SIGLAS

| | |
|---|---|
| AISI | American Iron and Steel Institute |
| API | American Petroleum Institute |
| ASME | American Society of Mechanical |
| ASTM | American Society for Testing and Materials |
| CSV | Comma-Separated Values |
| EI | Extração de Informações |
| FD | Folha de Dados |
| GPT | Generative Pre-trained Transformers |
| HTML | Hypertext Markup Language |
| ISO | International Organization for Standardization |
| LLM | Large Language Model |
| MSS | Manufacturers Standardization Society |
| NLTK | Natural Language Toolkit |
| NPT | National Pipe Thread |
| NIST | National Institute of Standards and Technology |
| PDF | Portable Document Format |
| PLN | Processamento de Linguagem Natural |
| PSI | Pound per Square Inch |
| RFQ | Request for Quotation |
| TXT | Text File |
| XLSX | Extensão de arquivo do Microsoft Excel. |

# LISTA DE FIGURAS

# LISTA DE GRÁFICOS

# SUMÁRIO

# 1. INTRODUÇÃO

Na dinâmica econômica atual do setor industrial, os orçamentos técnicos são cruciais, representando uma significativa fonte de receita para as empresas. Neste contexto, a agilidade e precisão na elaboração de propostas comerciais são vitais. Em um mercado altamente competitivo e tecnológico, onde diversas empresas disputam a atenção e o negócio dos clientes, a eficácia das propostas pode ser decisiva na captação e manutenção de relações comerciais. Este cenário destaca a importância de otimizar o processo de criação e análise de orçamentos técnicos, não apenas para melhorar a eficiência operacional, mas também como um elemento chave para o sucesso econômico e estratégico das empresas no mercado global.

Mesmo o foco sendo a triagem de orçamentos técnicos em que os objetos dos mesmos são válvulas industriais, precisamos contextualizar com a definição o que são as mesmas e o tamanho desse mercado.

## 1.1 Definição e tamanho estimado de mercado

Historicamente, a tecnologia tem impulsionado a evolução de diversos setores, incluindo o de design de válvulas, que remonta ao Império Romano, figura 1.1.

Figura 1- Válvula tipo Macho. Imagem cortesia do Ministério para Melhores Atividades Culturais, Bens Arqueológicos de Nápoles e Pompeia.

Válvulas industriais, figura 2, são dispositivos mecânicos utilizados para controlar, regular e direcionar o fluxo de líquidos, gases e sólidos em sistemas e processos industriais. Elas são fundamentais para garantir a operação segura e eficiente de equipamentos e instalações em diversos setores da indústria, desde a petroquímica até a alimentícia, figura 3.


Figura 2 - Tipos de válvulas industriais

Figura 3- Infraestrutura de Válvulas Industriais em Operação

Dada sua aplicação, as válvulas são projetadas para suportar altas pressões, temperaturas extremas e ambientes corrosivos, tornando-se essenciais em muitos processos industriais.

O mercado de válvulas industriais, avaliado em USD 59,2 bilhões em 2021, deve alcançar USD 95,07 bilhões até 2030, impulsionado pelo crescimento das atividades de infraestrutura. O gráfico 1 ilustra a evolução do mercado interno norte-americano na última década, refletindo a distribuição dos tipos de válvulas, comparável globalmente.

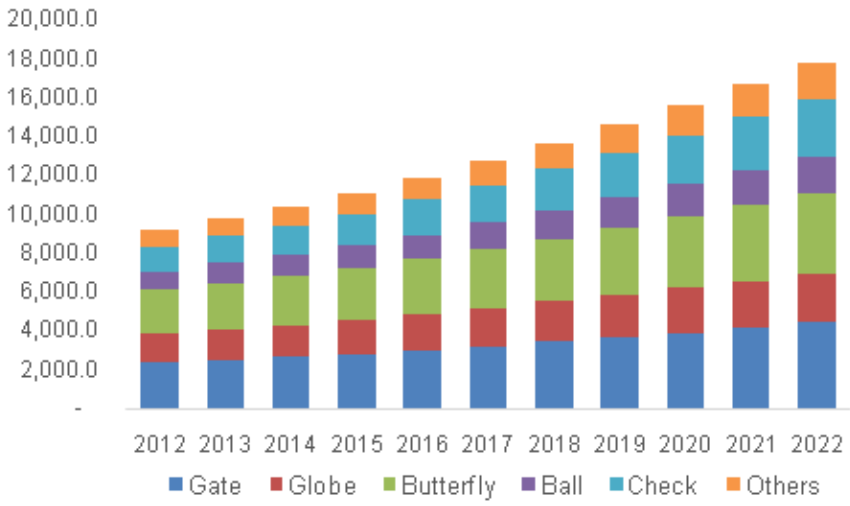



Gráfico 1 - Participação de mercado de válvulas industriais dos EUA por produto, 2012 - 2022 (em milhões de USD). Fonte: [Grand View Research] (https://www.grandviewresearch.com/industry-analysis/industrial-valves-market)

### 1.2 Cenário de vendas no mundo globalizado

No cenário atual, em um mundo globalizado, os processos comerciais enfrentam desafios crescentes. Com uma concorrência mais acirrada e clientes armados com uma abundância de informações, tornando-se mais críticos e exigentes (Kotler & Keller, 2016), as empresas são pressionadas a encontrar maneiras inovadoras de se sobressair e aprimorar seus processos. A necessidade de inovação é evidente em todas as etapas do ciclo empresarial, mas uma área que tem sido particularmente enfatizada é o ciclo de vendas (Ries, 2011). Dentro deste ciclo, o processo de elaboração de propostas e cotações é crucial, pois determina o primeiro ponto de compromisso formal com os clientes potenciais. Este trabalho de conclusão de curso mergulha profundamente neste aspecto, focando especificamente no nicho das válvulas industriais. Reconhecendo os desafios inerentes a este setor, o estudo propõe soluções avançadas através da integração de técnicas de aprendizado de máquina e Processamento de Linguagem Natural (PLN) (Goodfellow et al., 2016), visando otimizar e até mesmo revolucionar o processo tradicional.

### 1.3 Contextualização do Processo de Cotação e Proposta no Ciclo de Vendas

O processo de cotação e formalização de propostas não é apenas uma mera etapa dentro do ciclo de vendas. Ele é, na verdade, um momento estratégico que permite às empresas demonstrarem sua singularidade e aptidão para satisfazer as demandas específicas dos clientes (Kotler & Keller, 2016). Uma proposta bem articulada e estruturada pode ser determinante não só para a concretização de uma venda, mas também como um alicerce para estabelecer e nutrir relacionamentos comerciais duradouros (Cialdini, 2006). No entanto, em setores como o industrial, onde a diversidade de produtos e as complexidades técnicas são a norma, essa tarefa assume níveis adicionais de desafio e profundidade.

### 1.4 Desafios no Processo de Proposta e Cotação

A tarefa de criar propostas comerciais que equilibrem o rigor técnico com o apelo mercadológico representa um desafio contínuo e multifacetado para as equipes de vendas (Kotler & Keller, 2016). Estas propostas, que são fundamentais para estabelecer a primeira impressão e valor da empresa, frequentemente enfrentam obstáculos como a coleta manual de dados, o risco de incorporar informações imprecisas e a utilização de processos que podem estar desatualizados (Christensen, 1997). Diante da crescente demanda por eficiência, a automação desse processo torna-se essencial para otimizar as vendas e enriquecer a experiência do cliente. No setor industrial, é comum que departamentos de compras, juntamente com equipes de engenharia, encaminhem listas de equipamentos aos fabricantes. Contudo, a falta de especificações claras nessas listas amplia a responsabilidade dos fabricantes em garantir a conformidade, especialmente no segmento de válvulas industriais, onde a variedade de produtos e os orçamentos semiestruturados intensificam a complexidade da triagem, impactando diretamente na eficiência e nos custos operacionais. A figura 4 ilustra a complexidade embutida ao longo desse processo. A transição para a digitalização e automação, como observado no setor tecnológico-industrial, é vista como uma resposta estratégica para enfrentar desafios de mercado, como a demanda por produtos personalizados e a pressão por produção mais rápida e econômica (Narayanan et al., 2020).

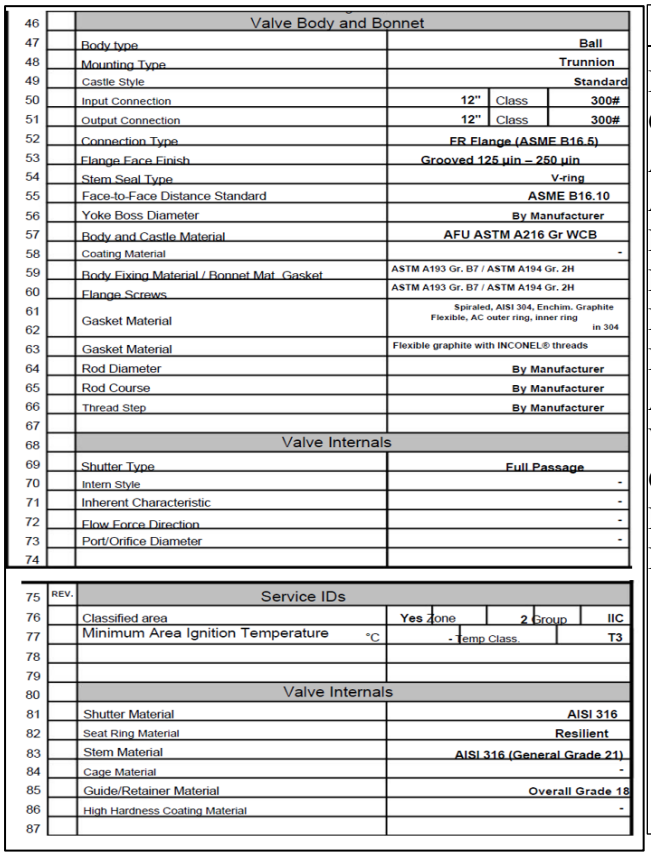

Figura 4 - Visão Geral da Complexidade no Processo de Triagem de Orçamentos e Propostas Técnicas

### 1.5 Objetivos da Pesquisa

Este estudo tem como objetivo principal a criação de um sistema automatizado que possa extrair, de forma eficiente, informações sobre válvulas a partir de documentos técnicos. Utilizando técnicas avançadas de PLN e aprendizado de máquina, busca-se otimizar o processo de triagem e classificação de propostas comerciais. A ideia é proporcionar às empresas do setor uma ferramenta que as auxilie a tornar o processo de vendas mais ágil, preciso e eficiente. Para ilustrar a aplicação prática deste sistema, apresentamos um exemplo de uma parte de uma Folha de Dados (FD) de uma válvula.

Nas duas figuras a seguir, observamos abordagens distintas na apresentação de dados técnicos. A Figura 5 ilustra uma Folha de Dados (FD) real, a qual, frequentemente, contém um excesso de informações para o projetista, ao mesmo tempo em que pode negligenciar ou omitir detalhes fundamentais, como normas de fabricação, de testes, ou outros aspectos específicos de grande relevância. Em contraste, a Figura 6 apresenta um texto conciso e informativo, cuidadosamente elaborado a partir das informações da primeira figura. Este texto destaca as informações essenciais, permitindo que o projetista trabalhe de maneira padronizada e segura.

Este exemplo ressalta as discrepâncias entre as informações disponíveis: na primeira figura, os dados estão dispersos e, muitas vezes, incompletos, enquanto a segunda figura oferece uma configuração ideal, enfatizando as informações vitais para a elaboração de uma proposta técnico-comercial eficaz. Esta comparação sublinha a capacidade do sistema proposto em identificar e preencher lacunas nos dados técnicos. Ao garantir a inclusão e a correta interpretação de todas as informações pertinentes ao processo, o sistema otimiza tanto a precisão quanto a rapidez na preparação de propostas comerciais. Este método proporciona um fluxo de trabalho mais eficiente e confiável, otimizando o desempenho e a eficácia dos projetistas.



| | | | | | |
|---|---|---|---|---|---|
| 46 | | Valve Body and Bonnet | | | |
| 47 | | Body type | Ball | | |
| 48 | | Mounting Type | Trunnion | | |
| 49 | | Castle Style | Standard | | |
| 50 | | Input Connection | 12" | Class | 300# |
| 51 | | Output Connection | 12" | Class | 300# |
| 52 | | Connection Type | FR Flange (ASME B16.5) | | |
| 53 | | Flange Face Finish | Grooved 125 µin – 250 µin | | |
| 54 | | Stem Seal Type | V-ring | | |
| 55 | | Face-to-Face Distance Standard | ASME B16.10 | | |
| 56 | | Yoke Boss Diameter | By Manufacturer | | |
| 57 | | Body and Castle Material | AFU ASTM A216 Gr WCB | | |
| 58 | | Coating Material | - | | |
| 59 | | Body Fixing Material / Bonnet Mat. Gasket | ASTM A193 Gr. B7 / ASTM A194 Gr. 2H | | |
| 60 | | Flange Screws | ASTM A193 Gr. B7 / ASTM A194 Gr. 2H | | |
| 61 | | Gasket Material | Spiraled, AISI 304, Enchim. Graphite Flexible, AC outer ring, inner ring in 304 | | |
| 62 | | | | | |
| 63 | | Gasket Material | Flexible graphite with INCONEL® threads | | |
| 64 | | Rod Diameter | By Manufacturer | | |
| 65 | | Rod Course | By Manufacturer | | |
| 66 | | Thread Step | By Manufacturer | | |
| 67 | | | | | |
| 68 | | Valve Internals | | | |
| 69 | | Shutter Type | Full Passage | | |
| 70 | | Intern Style | - | | |
| 71 | | Inherent Characteristic | - | | |
| 72 | | Flow Force Direction | - | | |
| 73 | | Port/Orifice Diameter | - | | |
| 74 | | | | | |
| 75 | REV. | Service IDs | | | |
| 76 | | Classified area | Yes Zone | 2 Group | IIC |
| 77 | | Minimum Area Ignition Temperature °C | - Temp Class | | T3 |
| 78 | | | | | |
| 79 | | | | | |
| 80 | | Valve Internals | | | |
| 81 | | Shutter Material | AISI 316 | | |
| 82 | | Seat Ring Material | Resilient | | |
| 83 | | Stem Material | AISI 316 (General Grade 21) | | |
| 84 | | Cage Material | - | | |
| 85 | | Guide/Retainer Material | Overall Grade 18 | | |
| 86 | | High Hardness Coating Material | - | | |
| 87 | | | | | |

Figura 5 - Folha de Dados Válvula Esfera

| Descritivo de ideal |
|---|
| Ball Valve Type; Diameter 12 Inches; Pressure Class 300psi; Body Material in Carbon Steel ASTM A216 Gr WCB; Ball Material in Stainless Steel AISI316; Stem Material in AISI316; Coating Material – Not Applicable; Resilient Seat; Face to Face ASME 16.10; Extremity Flanged RF ASME B16.5; Groove Raised Face 125µin - 250µin; Bolted Bonnet; Bolts ASTM A193 Gr B7 / Nuts ASTM A194 Gr 2h; Packing In Flexible Graphite With Inconel Threads. Additional Info: Construction API 6D; Passage: Full Bore; Trunnion Mounted; Actuation Manual Gearbox W/ Side Handwheel; Hydrostatic Test API598 |

Figura 6 - Descritivo Válvula Esfera

Neste estudo, a meta principal é desenvolver um sistema automatizado capaz de decifrar e extrair informações pertinentes sobre válvulas industriais de orçamentos, contendo dados e informações técnicas com precisão e eficiência. Com a aplicação de técnicas de ponta de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, objetiva-se otimizar o processo de triagem, categorização e análise de propostas comerciais. Este avanço não apenas simplificará o processo de avaliação para as empresas do setor, mas também garantirá uma maior acurácia e velocidade na tomada de decisões. Além disso, o sistema proposto visa aprimorar o ciclo de vendas, garantindo que as empresas possam estabelecer relações comerciais mais sólidas e duradouras, beneficiando-se de um processo otimizado e adaptado às demandas do cenário comercial contemporâneo. Por fim, este projeto busca consolidar uma ferramenta que seja um verdadeiro diferencial competitivo no mercado de válvulas industriais, oferecendo soluções ágeis e precisas às complexidades enfrentadas na elaboração e avaliação de propostas técnicas.


Figura 7 - O Confronto entre Inovação e Métodos Tradicionais: Navegando pela Burocracia no Ambiente Industrial. Fonte de criação: DALL.E

### 1.6 Significado do Estudo

A era da digitalização transformou a maneira como as empresas operam, tornando a inovação e a automação essenciais para se manterem competitivas (Schwab, 2016). Em um mercado globalizado, a rapidez e precisão nas transações comerciais determinam o sucesso ou fracasso de muitos negócios (Porter & Heppelmann, 2014). No contexto das propostas e cotações, onde cada minuto conta, a automação emerge não apenas como uma solução para melhorar a eficiência operacional, mas também como um meio de potencializar a satisfação do cliente (Brynjolfsson & McAfee, 2014).

Ao desenvolver e implementar um sistema que otimize a triagem de propostas, as empresas do setor industrial, particularmente aquelas que lidam com válvulas, podem significativamente reduzir erros, minimizar retrabalhos e acelerar seus ciclos de vendas (Chui et al., 2016). Esse estudo se alinha às necessidades do mercado, visando abordar lacunas existentes nas práticas atuais de triagem de propostas e cotações (Rüßmann et al., 2015).

Ao fazer isso, ele busca estabelecer um novo benchmark para a indústria de válvulas industriais, enfatizando a importância da integração entre tecnologia e práticas comerciais (Manyika et al., 2013). Além disso, ao abordar essa temática, o estudo oferece insights valiosos para a academia, expandindo o corpo de conhecimento em automação de processos de vendas e gestão de propostas (Bughin et al., 2017). A longo prazo, as descobertas deste estudo têm o potencial de moldar a maneira como a indústria aborda a gestão de propostas, estabelecendo novos padrões e práticas recomendadas que beneficiarão não apenas as empresas, mas toda a cadeia de valor do setor (Westerman et al., 2014).

# 2. FUNDAMENTACAO TEORICA

## 2.1 Processamento de Linguagem Natural (PLN) na Extração de Informações

O Processamento de Linguagem Natural (PLN) é uma disciplina que opera na confluência da linguística computacional e da inteligência artificial, com o intuito de dotar as máquinas da capacidade de compreender, interpretar e reagir à linguagem humana de forma eficiente e precisa. A extração de informações (EI), uma subárea significativa do PLN, dedica-se à identificação e extração de entidades e suas relações a partir de textos (Manning et al., 2008).

A aplicabilidade da EI é ampla, abrangendo desde o jornalismo, onde facilita a extração de nomes e entidades, até a biomedicina, contribuindo para o mapeamento de relações entre genes e doenças (Cohen & Hunter, 2008). Em um contexto de proliferação de dados não estruturados, a capacidade de extrair informações relevantes de grandes volumes de texto torna-se essencial. A crescente demanda por ferramentas capazes de transformar conjuntos de dados textuais em informações estruturadas reforça essa necessidade (Jurafsky & Martin, 2019).

Os avanços contínuos em técnicas de aprendizado de máquina e a disponibilidade de grandes conjuntos de dados têm impulsionado progressos notáveis na EI, resultando em sistemas mais precisos e robustos, capazes de processar e interpretar textos em diversos domínios e idiomas (Chowdhury, 2003).

Em resumo, a EI, como subdomínio do PLN, destaca-se no cenário tecnológico atual, servindo como uma ferramenta essencial para organizações e acadêmicos navegarem pelo vasto mar de dados não estruturados.

## 2.2 Estudos Anteriores sobre Análise de Texto para Documentos Técnicos

A análise textual de documentos técnicos surge como um campo de pesquisa de crescente importância no cenário científico e tecnológico contemporâneo. A literatura técnica e científica, caracterizada por sua complexidade e linguagem especializada, cresceu exponencialmente, demandando ferramentas e métodos capazes de extrair e sintetizar informações eficientemente (Liddy, 2001). Tais documentos abrigam um conhecimento valioso que, se adequadamente utilizado, pode impulsionar avanços significativos em diversas áreas, como medicina, engenharia e ciência da computação.

A crescente necessidade de processar esse vasto volume de informações tem fomentado o desenvolvimento de técnicas avançadas de PLN. Estas técnicas permitem não somente a extração de informações, mas também análises de sentimentos, categorização de documentos e sumarização automática (Hirschberg & Manning, 2015). Pesquisas anteriores demonstraram a eficácia de algoritmos de aprendizado de máquina na identificação de padrões e tendências em documentos técnicos, facilitando a descoberta de insights e suportando a tomada de decisões baseada em evidências (Sebastiani, 2002).

Os avanços contínuos em aprendizado de máquina e a disponibilidade de grandes conjuntos de dados têm impulsionado progressos notáveis na EI, resultando em sistemas mais

precisos e robustos, aptos a processar e interpretar textos em uma variedade de domínios e idiomas. Estes avanços incluem a disponibilidade de grandes conjuntos de dados de texto anotados, essenciais para treinar sistemas de EI precisos; o desenvolvimento de algoritmos de aprendizado de máquina mais avançados, capazes de identificar padrões complexos na linguagem humana; e a melhoria na disponibilidade de recursos computacionais, que permitem treinar e executar sistemas de EI de maneira mais rápida (Chowdhury, 2003).

Contudo, muitos estudos existentes, como o de Rizvi et al. (2018), assumem a necessidade de recursos externos, como ontologias, que podem não estar disponíveis no domínio atual. Esta dependência de ontologias ou de outros recursos externos pode limitar a aplicabilidade de tais sistemas em domínios altamente especializados, como o de válvulas industriais. A criação e manutenção de ontologias demandam um entendimento profundo e detalhado do domínio, além de ser um processo contínuo para assegurar sua atualização e relevância.

Neste estudo, a abordagem para a Extração de Informações (EI) no setor de válvulas industriais concentra-se no desenvolvimento de um dicionário técnico detalhado. Este recurso meticuloso é essencial para uma compreensão precisa da linguagem técnica especializada, garantindo a acurácia e relevância das informações extraídas neste contexto específico. Ao invés de depender de ontologias pré-existentes, que podem ser limitadas ou inacessíveis, este estudo opta por um método mais adaptável e focado no domínio específico. Esta abordagem detalhada assegura que cada termo técnico seja analisado com precisão, reforçando a aplicabilidade e a relevância dos resultados no campo das propostas técnicas de válvulas industriais.

# 3. METODOLOGIA

## 3.1 Contextualização

A metodologia adotada neste estudo visa aprimorar a triagem de orçamentos técnicos no setor de válvulas industriais. Para complementar a análise, conduziu-se uma pesquisa eletrônica, detalhada no Apêndice A, por meio de um formulário online no MS Forms, direcionada especificamente a profissionais envolvidos na elaboração diária de cotações e propostas técnicas em fabricantes e distribuidores de válvulas industriais. Participaram desta pesquisa entidades brasileiras, norte-americanas e chinesas, contribuindo com respostas a um conjunto de 9 perguntas estrategicamente formuladas para investigar o processo de recebimento, triagem e elaboração de propostas técnicas.

As questões aplicadas buscaram compreender a dinâmica do recebimento e manejo de propostas, identificar as extensões de arquivos predominantes, mapear as dificuldades mais comuns e quantificar o tempo médio investido nessas atividades. As respostas coletadas, que serão exibidas em detalhe nos apêndices deste documento, fornecem uma base sólida para a proposta de integrar técnicas de Aprendizado de Máquina, com o objetivo de potencializar a eficiência do processo. A disposição dos fabricantes e distribuidores em participar da pesquisa enriqueceu significativamente o estudo com perspectivas práticas. Os dados comparativos, que serão discutidos na seção de conclusão, endossam a aplicabilidade e a relevância da metodologia proposta para o campo em questão.

As solicitações de orçamento, conhecidas como RFQs (Request for Quotation), chegam em uma gama diversificada de formatos, incluindo PDF, XLSX, DOCX, TXT e até mesmo incorporadas no corpo de e-mails, HTML. Essa diversidade e a falta de padronização impõem a necessidade de uma abordagem de análise sistemática e meticulosa que, frequentemente, é onerosa em termos de tempo. No gráfico 2, ilustra proporção dos tipos de extensões de arquivos recebidos pelos usuários.

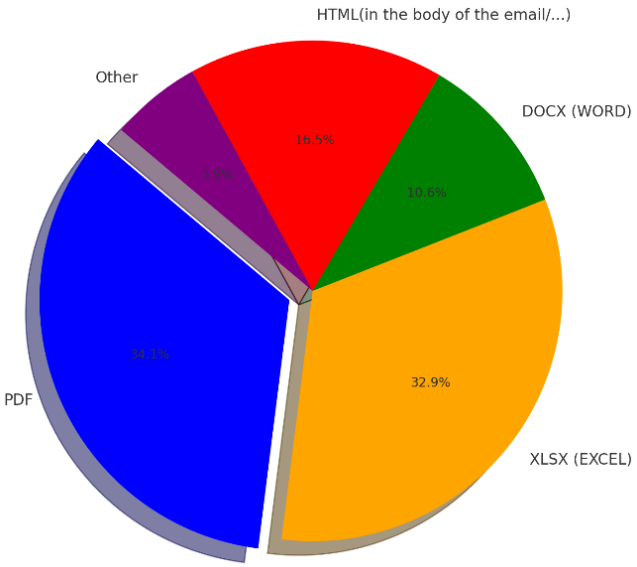



Gráfico 2 - Distribuição Percentual dos Formatos de Arquivos Recebidos por Email

É relevante destacar que, para a implementação do código computacional, optou-se pelo uso da linguagem Python, juntamente com suas bibliotecas especializadas e uma variedade de outras bibliotecas focadas no processamento do PLN. A relação completa dessas bibliotecas será vista a diante.

Mesmo o idioma desse trabalho sendo o Português, a língua Inglesa foi selecionada como o idioma base para a análise de propostas, não apenas devido à sua predominância no corpus técnico global, mas também porque as normas de construção de válvulas são majoritariamente estabelecidas por entidades como o API (American Petroleum Institute), ISO (International Organization for Standardization) e MSS (Manufacturers Standardization Society), ASME(American Society of Mechanical Engineers) que utilizam o inglês como idioma padrão em suas documentações e especificações. Além disso, a língua inglesa e o padrão utilizado pelos principais fabricantes, distribuidores e Traders de válvulas industriais.

## 3.2 Coleta e Pré-processamento dos Dados

O pré-processamento de dados é fundamental em qualquer estudo de Processamento de Linguagem Natural (PLN), preparando o terreno para análises mais detalhadas e sofisticadas. Singh (2023) ressalta a importância dessa fase preliminar, particularmente em situações em que os dados apresentam grande variação e complexidade estrutural. A pesquisa eletrônica complementar, realizada como parte deste estudo, fornece uma camada adicional de validação prática, endossando a demanda por um sistema de triagem de orçamentos mais eficaz, tal como relatado pelos profissionais que lidam com esses desafios rotineiramente.

A coleta de dados envolve a importação de documentos do repositório específico mencionado anteriormente. Uma vez coletados, os dados passam por uma fase de pré-processamento. Esta fase é essencial para garantir que os dados estejam em um formato adequado para análise. O pré-processamento pode envolver a remoção de ruídos, a correção de erros e a conversão de formatos, entre outras tarefas. O objetivo é garantir que os dados estejam prontos para análise.

A metodologia sugerida propõe que os orçamentos sejam inicialmente importados de um repositório designado pelo usuário. O sistema desenvolvido, ao identificar a extensão do arquivo, converte-o automaticamente para um formato de texto simples, o que otimiza o processamento subsequente. Após a conversão, os textos são submetidos a uma limpeza, através de processos como tokenização e lematização, que segmentam o texto em unidades básicas, simplificando a identificação e a classificação das informações relevantes.

## 3.3 Processamento de Arquivos no PLN

### 3.3.1 Importação de Arquivos: PDF, XLSX e TXT

A importação de arquivos é o primeiro passo crucial na análise de dados, especialmente quando se lida com documentos técnicos em diferentes formatos. Arquivos em formatos PDF, TXT e XLSX são comuns em ambientes corporativos e acadêmicos, e cada um apresenta seus

próprios desafios e peculiaridades. O formato PDF, por exemplo, é amplamente utilizado para a publicação de documentos técnicos devido à sua capacidade de preservar a formatação original. Para o estudo em questão, o PDF é a extensão de arquivos de orçamentos mais recebidas pelo mercado de válvulas industriais, como corroborado pela pesquisa, cujo resultado é ilustrado na gráfico 3.1. Por outro lado, arquivos TXT são mais simples e contêm apenas texto, tornando a extração de informações mais direta. Já os arquivos XLSX, associados ao Microsoft Excel, são frequentemente usados para armazenar dados tabulares, e sua manipulação requer técnicas específicas para acessar e interpretar as informações contidas nas planilhas (Stein, 2022).

A manipulação de arquivos em formatos como PDF, XLSX e TXT no contexto do Processamento de Linguagem Natural (PLN) é uma tarefa fundamental para a extração e análise de dados. Cada um desses formatos apresenta desafios e características únicas que requerem abordagens específicas. A Tabela 3.1, recebida em formato PDF, ilustra um orçamento real. A extração de texto de arquivos PDF pode ser desafiadora devido à sua natureza binária e à possibilidade de conter elementos gráficos e multimídia. No entanto, técnicas avançadas têm sido desenvolvidas para superar esses desafios. Um estudo de Kuckartz e Rädiker (2019) discute métodos para codificar e analisar textos em arquivos PDF, destacando a importância de ferramentas especializadas para garantir a precisão da extração de dados.

ALUMAR — HATCH — Maranhão Aluminum Consortium

FEL 3 - BASIC ENGINEERING
REFINERY PRODUCTION DEBOTTLENECKING PHASE #2
ANNEX 1 - WORKSHEET OF QUANTITATIVES
PACOTE PP004

Nº EDM: SLU-M-002803_A1 | PÁGINA: 1/1
Nº HATCH: H357751-PP004-250-242-0001_A1 | REV.: 3A

| Rev. | Item | Área | Sub-área | Pacote | TAG (Tipo) | DESCRIPTION | Unidade | QTY | UNIT PRICE | TOTAL PRICE | REMARKS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | | | REFINERY PRODUCTION DEBOTTLENECKING PHASE #2 | | | | | |
| | 1.1 | 2 | | | | RED AREA | | | | | |
| | 1.1.1 | 2 | 025 | PP004 | | Bauxite Grinding | | | | | |
| | 1.1.1.1 | 2 | 025 | PP004 | | VALVES | | | | | |
| 3A | 1.1.1.1.1 | 2 | 025 | PP004 | V111.1 | V111.1 - 3/4" - BALL VALVE, CLASS 300#, MANUFACTURER STANDARD, NPT THREADED, BODY IN ASTM A216 GR. WCB, BALL IN AISI 316 | pç | 26,00 | | | |
| 3A | 1.1.1.1.2 | 2 | 025 | PP004 | V111.1 | V111.1 - 2" - BALL VALVE, CLASS 300#, MANUFACTURER STANDARD, NPT THREADED, BODY IN ASTM A216 GR. WCB, BALL IN AISI 316 | pç | 28,00 | | | |
| 3A | 1.1.1.1.3 | 2 | 025 | PP004 | V50.1 | V50.1 - 2" - PLUG VALVE, B16.10, CLASS 150#, RAISED FACE FLANGE, SHORT PATTERN, BODY IN A216 Gr. WCB, SERRATED CONCENTRIC ACCORDING TO MSS SP6 | pç | 14,00 | | | |
| 3A | 1.1.1.1.4 | 2 | 025 | PP004 | V50.2 | V50.2 - 2" - PLUG VALVE, ASTM A216 GR. WCB, FLG. FR, MSS SP6, CLASS 150#, ASME B16.5, ASME B16.10, ASME B16.34 | pç | 25,00 | | | |
| 3A | 1.1.1.1.5 | 2 | 025 | PP004 | V52.1 | V52.1 - 2" - PLUG VALVE, B16.10, CLASS 150#, FLG. FR, SHORT PATTERN, BODY IN ASTM A395, REV. EM PFA CONF. ASTM F781, PLUG IN STAINLESS STEEL, MSS SP6 | pç | 8,00 | | | |
| 3A | 1.1.1.1.6 | 2 | 025 | PP004 | AP114 | AP114 - 4" - LINE BLIND VALVE (STACEY) | pç | 22,00 | | | |
| 3A | 1.1.1.1.7 | 2 | 025 | PP004 | V51.2 | V51.2 - 6" - PLUG VALVE, ASTM A216 GR. WCB, FLG. FR, MSS SP6, CLASS 150#, ASME B16.5, ASME B 16.10, ASME B16.34 | pç | 1,00 | | | |
| | | | | | | | | | | | |
| | 1.1.2 | 2 | 025A | PP004 | | Bauxite Mud Mixture | | | | | |
| | 1.1.2.1 | 2 | 025A | PP004 | | VALVES | | | | | |
| 3A | 1.1.2.1.1 | 2 | 025A | PP004 | V111.1 | V111.1 - 3/4" - BALL VALVE, CLASS 300#, MANUFACTURER STANDARD, NPT THREADED, BODY IN ASTM A216 GR. WCB, BALL IN AISI 316 | pç | 12,00 | | | |
| 3A | 1.1.2.1.2 | 2 | 025A | PP004 | V111.1 | V111.1 - 2" - BALL VALVE, CLASS 300#, MANUFACTURER STANDARD, NPT THREADED, BODY IN ASTM A216 GR. WCB, BALL IN AISI 316 | pç | 5,00 | | | |
| 3A | 1.1.2.1.3 | 2 | 025A | PP004 | V50.1 | V50.1 - 2" - PLUG VALVE, B16.10, CLASS 150#, RAISED FACE FLANGE, SHORT PATTERN, BODY IN A216 Gr. WCB, SERRATED CONCENTRIC ACCORDING TO MSS SP6 | pç | 5,00 | | | |
| 3A | 1.1.2.1.4 | 2 | 025A | PP004 | V50.2 | V50.2 - 2" - PLUG VALVE, ASTM A216 GR. WCB, FLG. FR, MSS SP6, CLASS 150#, ASME B16.5, ASME B16.10, ASME B16.34 | pç | 3,00 | | | |
| 3A | 1.1.2.1.5 | 2 | 025A | PP004 | V54.1 | V54.1 - 2" - PLUG VALVE, ASTM A216 GR. WCB, NPT THREADED, B1.20.1, CLASS 150#, B16.34 | pç | 1,00 | | | |
| 3A | 1.1.2.1.6 | 2 | 025A | PP004 | V51.2 | V51.2 - 6" - PLUG VALVE, B16.10, CLASS 150#, RAISED FACE FLANGE, SHORT PATTERN, BODY IN A216 Gr. WCB, SERRATED CONCENTRIC ACCORDING TO MSS SP6 | pç | 2,00 | | | |

Figura 8 - Planilha de Quantitativos e Especificações de Válvulas

Por fim, os arquivos XLSX, que são formatos de planilha do Microsoft Excel, contêm dados tabulares que podem ser de interesse para análises de PLN. A manipulação desses arquivos requer a extração de dados de células, linhas e colunas específicas, e a conversão desses dados em um formato que possa ser processado por algoritmos de PLN. Os arquivos TXT, por outro lado, são formatos de texto simples e, portanto, mais fáceis de serem manipulados. Devido a essa facilidade que os arquivos desse estudo são inicialmente transformados em arquivo de texto, pois são usados para armazenar dados brutos ou transcrições. A análise de arquivos TXT foi abordada por diversos autores, como é o caso do trabalho de 2009 que discute o processamento de arquivos de texto e destaca que "o texto está em toda parte" (Apress, 2009). As técnicas e ferramentas desenvolvidas para lidar com esses formatos são cruciais para a extração eficaz de informações e a realização de análises aprofundadas. A seguir, encontra-se parte inicial do código, escrito em Python, que faz a interação com o usuário e faz a conversão da extensão de arquivo recebida para extensão de texto. Abaixo, na figura 7, parte do Código para transformar inicialmente extensões PDF e EXCEL em arquivo de texto.

```
import re
import nltk
import pandas as pd
import csv
from io import String IO
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser

# Baixar o pacote NLTK necessário
nltk.download('punkt')

def ler_pdf(caminho_arquivo):
    output_string = StringIO()
    with open(caminho_arquivo, 'rb') as f:
        parser = PDFParser(f)
        doc = PDFDocument(parser)
        rsrcmgr = PDFResourceManager()
        dispositivo = TextConverter(rsrcmgr, output_string, laparams=LAParams())
        interpretador = PDFPageInterpreter(rsrcmgr, dispositivo)
        for pagina in PDFPage.create_pages(doc):
            interpretador.process_page(pagina)
    return output_string.getvalue()

# Solicitar entrada do usuário
escolha = input("Escolha a fonte de dados (Digite 'excel', 'txt' ou 'pdf'): ").strip().lower()
if escolha == 'excel':
    # Carregar o conjunto de dados de descrição de válvulas de um arquivo Excel
    arquivo_excel = "caminho_para_arquivo_excel.xlsx"
    df = pd.read_excel(arquivo_excel)
    texto = ' '.join(map(str, df[df.columns[0]].tolist())).lower()
elif escolha == 'txt':
    try:
        with open("caminho_para_arquivo_texto.txt", "r", encoding="utf-16") as f:
            texto = f.read().lower()
    except UnicodeDecodeError:
        print("Erro ao ler o arquivo com codificação utf-16. Por favor, verifique a codificação do arquivo.")
        exit()
elif escolha == 'pdf':
    arquivo_pdf = caminho_para_arquivo_texto.pdf"
    texto = ler_pdf(arquivo_pdf).lower()
else:
    print("Escolha inválida. Por favor, digite 'Excel', 'txt' ou 'pdf'.")
    exit()
    ...
    (Continua)
```

Figura 9 - Script de Código em Python para Triagem de Válvulas Industriais

Como resultado a seguir, a saída é um texto em caixa baixa, ainda não estruturado. Apenas, pronto para a próxima fase, que será a de preparação para parametrizar, limpar e classificar as sentenças inicialmente.

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!

Choose the data source (Enter 'excel', 'txt', or 'pdf'): pdf

bauxite mud mixture

valves
ball valve, class 300#, manufacturer standard, npt threaded, body in
astm a216 gr. wcb, ball in aisi 316
v111.1 - 2" - <> ball valve, class 300#, manufacturer standard, npt threaded, body in astm
a216 gr. wcb, ball in aisi 316
v50.1 - 2" - <> plug valve, b16.10, class 150#, raised face flange, short pattern, body in a216
gr. wcb, serrated concentric according to mss sp6
v50.2 - 2" - <> plug valve, astm a216 gr. wcb, flg. fr, mss sp6, class 150#, asme b16.5, asme
b16.10, asme b16.34
v52.1 - 2" - <> plug valve, b16.10, class 150#, flg. fr, short pattern, body in astm a395, rev.
em pfa conf. astm f781, plug in stainless steel, mss sp6
ap114 - 4" - line blind valve  (stacey)
v51.2 - 6" - <> plug valve, astm a216 gr. wcb, flg. fr, mss sp6, class 150#, asme b16.5, asme b
16.10, asme b16.34

bauxite mud mixture

valves
v111.1 - 3/4" - <> ball valve, class 300#, manufacturer standard, npt threaded, body in
astm a216 gr. wcb, ball in aisi 316
v111.1 - 2" - <> ball valve, class 300#, manufacturer standard, npt threaded, body in astm
a216 gr. wcb, ball in aisi 316
v50.1 - 2" - <> plug valve, b16.10, class 150#, raised face flange, short pattern, body in
a216 gr. wcb, serrated concentric according to mss sp6
v50.2 - 2" - <> plug valve, astm a216 gr. wcb, flg. fr, mss sp6, class 150#, asme b16.5, asme
b16.10, asme b16.34
v54.1 - 2" - <> plug valve, astm a216 gr. wcb, npt threaded, b1.20.1, class 150#, b16.34
v51.2 - 6" - <> plug valve, b16.10, class 150#, raised face flange, short pattern, body in
a216 gr. wcb, serrated concentric according to mss sp6
```

Figura 10 - Exemplo de Texto Extraído e Pré-processado para Análise em PLN

### 3.3.2 Processamento de Texto em PLN

No Processamento de Linguagem Natural (PLN), após importar e realizar o pré-processamento dos arquivos textuais, a tokenização é o próximo passo crítico. Esta fase divide o texto em tokens - palavras, frases ou parágrafos - que são as unidades fundamentais para a análise linguística. A tokenização de sentenças é particularmente importante, pois desagrega o texto em componentes individuais, facilitando a identificação de padrões e relações únicas no material analisado. Ferramentas como NLTK e SpaCy são imprescindíveis nesse processo, permitindo uma tokenização precisa e robusta, essencial para o tratamento eficiente de textos (Muse, 2023).

### 3.3.3 Aplicação da Tokenização em Textos Técnicos

A tokenização, essencial para segmentar textos técnicos em tokens de acordo com as necessidades analíticas, é exemplificada pela decomposição de uma descrição técnica em um orçamento de válvulas industriais. A frase:

*"Válvula de esfera, montagem tipo trunnion, construção conforme API 6D, corpo em aço carbono ASTM A216 Gr. WCB, esfera em aço inoxidável AISI 316, extremidades flangeadas tipo face ressalto com ranhuras concêntricas de acordo com MSS SP-6, flange seguindo ASME B16.5, passagem plena, diâmetro nominal de 200 mm, classe de pressão de 300 psi, sede primária resiliente em RPTFE, acionamento por caixa de engrenagens e volante lateral, atendendo a requisitos especiais para baixas emissões fugitivas conforme API 641 ou ISO 15848-1."*

é fragmentada em elementos como "montagem tipo trunnion" e "classe de pressão de 300 psi".

A tokenização eficiente, realizada com o auxílio das mencionadas ferramentas de PLN, destaca as especificações críticas para uma interpretação técnica acurada. Além disso, é vital reconhecer e abordar os desafios linguísticos encontrados em textos técnicos, um aspecto ressaltado por Goodfellow, Bengio e Courville (2016), onde técnicas avançadas de Deep Learning apresentam novos métodos para uma tokenização que contribui para a automação e a precisão na análise de propostas técnicas complexas.

### 3.3.4 Limpeza do Texto: Remoção de Palavras Irrelevantes e Lematização

As etapas de limpeza e pré-processamento de texto são fundamentais em projetos de PLN, pois preparam o texto para análise posterior, removendo ruídos e consolidando informações essenciais. A remoção de palavras irrelevantes, ou 'stop words', é crucial para evitar distorções nas análises e economizar recursos computacionais. Em textos técnicos do domínio industrial, como os relacionados `as válvulas, esse processo é particularmente importante para focar nos termos técnicos, removendo palavras comuns como 'e', 'ou' e 'mas'.

A lematização, o processo de reduzir as palavras a sua forma base ou canônica, é uma prática recomendada para manter a consistência na análise dos termos técnicos. Para este estudo, adotou-se o WordNetLemmatizer do NLTK, que é baseado no WordNet, um extenso banco de dados léxico para o inglês (Fellbaum, 1998). A lematização é fundamental para tratar variações do mesmo termo técnico, como 'válvula', 'válvulas', 'valvulado', ' valves' ou 'valve', permitindo que sejam agrupadas como uma única entidade lexical. Este procedimento facilita o tratamento de termos como 'flange', 'flanges', 'flangeamento', e 'acionamento', 'acionamentos', entre outros, garantindo que as variações sejam reconhecidas e analisadas corretamente.

A combinação de remoção de 'stop words' com a lematização assegura a limpeza e padronização do texto, tornando-o adequado para análises detalhadas. Este passo é crucial para garantir que descrições técnicas, como as de diferentes tipos de válvulas, sejam corretamente identificadas e interpretadas. A eficácia do WordNetLemmatizer no tratamento de termos técnicos é complementada por uma discussão sobre ferramentas alternativas para diferentes idiomas, inspirada em Jurafsky e Martin (2019), que abordam a diversidade linguística em PLN.

### 3.3.5 Delimitação de sentenças

Na etapa subsequente à limpeza textual, o sistema executa um filtro automático, selecionando apenas sentenças que mencionam válvulas ou suas variantes. Sentenças sem esses termos são descartadas, refinando o conteúdo para análise especializada. Esta seleção criteriosa destaca trechos técnicos específicos, delimitando-os de forma precisa para facilitar a identificação e interpretação detalhada de cada aspecto técnico relacionado a válvulas industriais.

O procedimento de delimitação enfatiza os termos que comumente iniciam descrições técnicas, como os tipos de válvulas e suas formas lexicais, e define o término de cada segmento descritivo. As figuras anexas ilustram esse processo em etapas: a Figura 9 mostra um trecho da lista técnica original; a Figura 10 revela o texto após a limpeza e padronização; e a Figura 11 destaca as descrições técnicas já delimitadas, resultando em sentenças estruturadas e prontas para análises aprofundadas. Essa metodologia assegura que, ao final, o texto consolidado se concentre exclusivamente em sentenças que contêm descrições relevantes de válvulas, como demonstrado visualmente pelas figuras mencionadas.

| | | | |
|---|---|---|---|
| PP004 | V111.1 | V111.1 - 2" - BALL VALVE, CLASS 300#, MANUFACTURER STANDARD, NPT THREADED, BODY IN ASTM A216 GR. WCB, BALL IN AISI 316 | pç |
| PP004 | V50.1 | V50.1 - 2" - PLUG VALVE, B16.10, CLASS 150#, RAISED FACE FLANGE, SHORT PATTERN, BODY IN A216 Gr. WCB, SERRATED CONCENTRIC ACCORDING TO MSS SP6 | pç |

Figura 11 - Texto original recebido

v111.1 - 2" - <> ball valve, class 300#, manufacturer standard, npt threaded, body in astm a216 gr. wcb, ball in aisi 316
v50.1 - 2" - <> plug valve, b16.10, class 150#, raised face flange, short pattern, body in a216 gr. wcb, serrated concentric according to mss sp6

Figura 12 - Texto delimitado em sentenças com válvulas e limpo

| | |
|---|---|
| Sentence_3 | ball valve, class 300#, manufacturer standard, npt threaded, body in astm a216 gr. wcb, ball in aisi 316 v50.1 - 2" - |
| Sentence_4 | plug valve, b16.10, class 150#, raised face flange, short pattern, body in a216 gr. wcb, serrated concentric according to mss sp6 v50.2 - 2" - |

Figura 13 - Sentenças apenas contendo válvulas, em lista estruturada, salva em arquivo de texto tipo csv

### 3.4 Extração de Informações

A fase de extração de informações é um componente crítico na análise de documentos técnicos que abordam válvulas industriais. Utilizando um léxico técnico cuidadosamente desenvolvido e técnicas avançadas de Processamento de Linguagem Natural (PLN), esta fase se concentra em identificar, filtrar e categorizar informações relevantes. O objetivo é estruturar os dados para uma análise consistente e precisa.

#### 3.4.1 Identificação de Sentenças que Contenham a Palavra "VÁLVULA"

Nesta etapa, o foco é identificar sentenças contendo a palavra "válvula" ou suas variações. Para isso, utilizamos expressões regulares para uma busca precisa e eficaz no texto.

Exemplo de Código Python para esse trecho:

```
# Este código busca por ocorrências da palavra "válvula" como uma palavra independente nos textos, garantindo que apenas as frases relevantes sejam selecionadas para análise.

frases_valvulas = [frase for frase in frases_lematizadas if re.search(r'\bválvula\b', frase)]
```

#### 3.4.2 Filtragem Inicial Focada em Válvulas Industriais

A seguir, implementamos um filtro para isolar sentenças específicas relacionadas a válvulas industriais. Essa seleção se baseia em técnicas de PLN para diferenciar informações relevantes de conteúdos não relacionados.

Exemplo de Código Python:

```
#Neste exemplo, a expressão regular identifica frases que contêm tipos específicos de válvulas, assegurando a captura de informações pertinentes.

padrao = r"(válvula gaveta.*?|válvula esfera.*?|válvula globo.*?|válvula retenção.*?|válvula macho.*?)(?= válvula gaveta| válvula esfera| válvula globo| válvula retenção| válvula macho|$)"

frases_valvulas = re.findall(padrao, texto, re.DOTALL)
```

#### 3.4.3 Categorização de Sentenças com Base nos Tipos de Válvulas

Após filtrar as informações relevantes, procedemos com a categorização das sentenças com base nos tipos de válvulas identificados.

Exemplo de Código Python:

```
#Aqui, o código realiza a substituição dos tipos de válvulas no texto para uniformizar a formatação, facilitando a categorização subsequente.

tipos_valvulas = ["válvula gaveta", "válvula esfera", "válvula globo", "válvula retenção", "válvula macho"]

for valvula in tipos_valvulas:

    texto = re.sub(valvula, " " + valvula, texto)
```

### 3.4.4 Extração de Tamanhos Nominais, Classes de Pressão e Características Principais

Esta etapa dedica-se à extração de detalhes técnicos como tamanhos nominais e classes de pressão das válvulas.

Exemplo de Código Python:

```
# Este trecho do código identifica e extrai informações sobre os tamanhos nominais das válvulas, com base no vocabulário predefinido.

for termo in vocabulario["nominal_size"]:

  if termo in frase:

    palavras_vocabulario[frase].append(termo)
```

### 3.4.5 Reconhecimento de Outros Termos Relevantes

Finalmente, o sistema identifica outros termos técnicos essenciais para uma análise completa dos documentos.

Exemplo de Código Python:

```
#Este código busca por termos adicionais no texto, ampliando a análise para incluir componentes e especificações técnicas relacionadas às válvulas.

vocabulary = {"other_terms": ["body", "cover", "ball", "stem", "seat", "ring", "bolt", "flange", "gasket", "packing"]}

for termo in vocabulary["other_terms"]:

  if termo in frase:

    palavras_vocabulario[frase].append(termo)
```

Em síntese, esta abordagem detalhada para extração de informações garante uma análise precisa e abrangente de documentos técnicos, focada nas válvulas industriais e suas especificações. Este processo proporciona uma base sólida para a compreensão detalhada dos equipamentos.

## 4. IMPLEMENTAÇÃO

### 4.1 Configuração das Bibliotecas e Suas Aplicações Específicas

Nesta etapa do projeto, diversas bibliotecas Python são configuradas, cada uma desempenhando um papel essencial:

- **re:** Esta biblioteca é empregada para realizar expressões regulares, um aspecto crucial na identificação de padrões textuais específicos nos orçamentos, como os termos técnicos associados a válvulas industriais. A capacidade de filtrar e isolar informações relevantes através de expressões regulares é fundamental para a precisão analítica do projeto.
- **Natural Language Toolkit (nltk):** O NLTK é uma ferramenta essencial no processamento de linguagem natural (PLN), empregada para tarefas como a tokenização e a lematização dos textos dos orçamentos. Estas funcionalidades são imprescindíveis para a análise linguística, permitindo uma compreensão mais profunda da estrutura e do significado do texto.
- **pandas:** A utilização desta biblioteca concentra-se na organização dos dados extraídos em estruturas tabulares, conhecidas como DataFrames. O pandas simplifica a manipulação, análise e visualização dos dados, tornando-os mais acessíveis para análises subsequentes.
- **csv:** Esta biblioteca é utilizada para facilitar as operações de leitura e escrita em arquivos no formato CSV. Sua importância reside na capacidade de gerenciar eficientemente grandes conjuntos de dados, uma necessidade comum no contexto deste projeto.
- **io.StringIO:** Utilizado para operações de leitura e escrita de strings em memória, esta ferramenta otimiza o processamento de texto. Sua aplicação é particularmente relevante na manipulação de textos extraídos de arquivos PDF.
- **pdfminer:** Componentes específicos desta biblioteca são utilizados para a extração eficiente de texto de arquivos PDF, um formato frequentemente encontrado nos orçamentos recebidos. Esta funcionalidade é vital para a conversão de dados de formato PDF para um formato analisável.
- **WordNetLemmatizer e word_tokenize (nltk):** Estas ferramentas do NLTK são empregadas, respectivamente, para lematizar e tokenizar o texto. A lematização reduz as palavras às suas formas base, enquanto a tokenização segmenta o texto em unidades menores, facilitando assim a análise linguística detalhada.
- **collections.defaultdict:** Esta estrutura de dados é utilizada para organizar as informações extraídas em dicionários que armazenam listas de itens, facilitando a categorização e o acesso organizado aos dados.

Cada biblioteca contribui de forma única e significativa para a eficácia do projeto, permitindo uma análise rigorosa e sistemática dos dados coletados nos orçamentos de válvulas industriais.

### 4.3 Integração com PDFMiner e Pandas

A integração com PDFMiner e Pandas é uma etapa crucial na automação do processo de triagem e análise de propostas técnicas de válvulas industriais, como discutido em no estudo. A utilização do PDFMiner é estratégica para a conversão eficiente de arquivos PDF, que são comuns neste contexto, em texto legível e processável. O PDFMiner atua desmembrando o conteúdo dos PDFs, página por página, e transformando-os em texto estruturado. Este processo é vital, considerando a prevalência de orçamentos e especificações técnicas recebidas nesse formato.

Em paralelo, a adoção do Pandas, uma biblioteca Python para manipulação de dados, permite alocar os dados extraídos em DataFrames, que são estruturas tabulares. Esta abordagem facilita significativamente a manipulação e análise posterior dos dados. DataFrames oferecem uma representação intuitiva e acessível de dados, essencial para tarefas como filtragem, classificação e análise de grandes conjuntos de informações técnicas.

### 4.4 Validação e Testes

A fase de validação e testes constitui um elemento crucial para aferir a eficácia e a acurácia do algoritmo desenvolvido. Os testes foram conduzidos em pequena escala, envolvendo a análise de cerca de 15 orçamentos reais obtidos entre o final de setembro e o final de outubro de 2023. Estes orçamentos consistiam em listas de 10 a 15 itens diversificados, predominantemente no formato PDF e em Excel. A seleção desta amostra justifica-se por sua capacidade de corroborar os resultados indicados no formulário do Anexo A, que foi preenchido por profissionais da área de orçamentos, proporcionando uma validação contextualizada e relevante dos resultados do algoritmo.

Os testes consistiram na inserção das listas de orçamentos, uma de cada vez, evidenciando ainda a necessidade de um acompanhamento humano no processo de geração das propostas. A intervenção humana é necessária para garantir a integridade e a exatidão dos dados finais, apesar do sistema ter demonstrado uma melhora muito significativa na velocidade de processamento do processo.

Foi observado que o sistema apresentou algumas pendências que precisam ser aperfeiçoadas; contudo, o comportamento geral do algoritmo atendeu às expectativas. As pendências encontradas são principalmente refinamentos necessários para lidar com itens de descrições complexas ou formatos de dados não convencionais.

Embora seja necessária uma checagem humana na etapa de inserção dos dados, o balanço entre a aceleração do processamento e a precisão dos resultados é substancialmente positivo. O operador precisa realizar apenas uma checagem final dos itens e suas quantidades, o que, dada a eficiência do algoritmo, reduz-se mais a uma confirmação dos resultados do que a uma reelaboração integral da proposta.

Portanto, os testes confirmaram a robustez do sistema em um contexto de aplicação real e validaram o alinhamento das etapas de desenvolvimento e implementação do algoritmo com os objetivos do projeto. A solução apresentada, apesar das melhorias necessárias, demonstra ser uma ferramenta valiosa na análise e triagem de orçamentos de válvulas industriais, proporcionando uma base firme para futura otimização e aplicação em uma escala mais ampla.

# 5. RESULTADOS E DISCUSSÃO

## 5.1 Visão Geral das Informações de Válvulas Extraídas

A aplicação do sistema de aprendizado de máquina para análise de propostas de válvulas industriais demonstrou eficácia notável. Este sistema possibilitou a extração de um espectro abrangente de dados, que inclui, além de outros elementos, a identificação de tipos de válvulas, suas medidas, classes de pressão e materiais constituintes. Através deste processo automatizado, elaborou-se um inventário detalhado das especificações técnicas, fundamental para a tomada de decisões informadas e ágeis no contexto comercial.

A avaliação dos resultados englobou duas dimensões principais:

- Avaliação Quantitativa: Foi realizada uma análise quanto ao volume e à variedade dos tipos de válvulas identificadas no conjunto de dados. Esta etapa quantitativa possibilitou uma avaliação crítica da capacidade do sistema em processar e classificar uma gama diversificada de válvulas, destacando a eficiência do algoritmo em manejar e organizar de informações.
- Avaliação Qualitativa: Esta avaliação concentrou-se na exatidão e relevância das informações extraídas, assegurando conformidade com as normas técnicas estabelecidas pelas entidades como API, ASME e ISO. A análise incluiu a verificação da precisão dos dados, com especial atenção à correspondência entre as características das válvulas e suas especificações técnicas. Exemplos específicos de casos em que o sistema demonstrou alta precisão foram analisados para ilustrar a eficácia qualitativa do sistema.

Estes resultados indicam que, apesar de algumas áreas ainda requererem melhorias, como por exemplo a identificação das quantidades para cada item que normalmente não estão contidos no mesmo descritivo, a metodologia automatizada oferece eficiência significativa no processo de triagem de orçamentos. O equilíbrio entre a rapidez de processamento e a acurácia dos dados reflete um progresso substancial na otimização dos procedimentos de vendas e refletirá positivamente na satisfação das necessidades dos clientes, sugerindo grande potencial para implementações futuras em uma escala expandida.

## 5.1 Etapas de Análise e Padronização de Termos Técnicos

A eficácia do método proposto é evidenciada nas descrições a seguir. Inicialmente, dois descritivos provenientes da Tabela 3.1, que representam casos reais, são processados preliminarmente. A Figura 3.5 ilustra este processo, destacando a extração, purificação e organização das sentenças, as quais constituem a fundação para a análise conclusiva.

A análise final é estruturada em duas etapas críticas:

- A primeira etapa focaliza no texto das sentenças originárias da fase preliminar, como representado na Figura 3.5. O sistema procede com a identificação de termos técnicos a partir de um vocabulário extensivo predefinido. Nesta fase, são reconhecidos os termos pertinentes no texto e são formuladas descrições básicas dos elementos encontrados. Termos não localizados são excluídos sem menção adicional, mantendo a integridade do texto.

| Label | Sentence |
|---|---|
| Sentence_nb | full description valve type |
| Sentence_3 | ball valve, nominal size 2inches, pressure class 300psi, body material in astm a216 gr. wcb, ball material aisi 316, construction manufacturer standard, end connections npt threaded, other terms v50.1 |
| Sentence_4 | plug valve, nominal size 2inches, pressure class 150psi, body material in astm a216 gr. wcb, end connection raised face flange , standards b16.10 mss sp6, other terms serrated concentric v50.2 |

Figura 14 - Descrição básica

- A segunda etapa envolve a análise comparativa das informações coletadas para cada tipo de válvula em relação a um repositório complementar que abarca normativas específicas aplicáveis a cada modelo. Discrepâncias identificadas ou ausências de termos importantes são automaticamente integradas pelo sistema, que, adicionalmente, uniformiza a configuração dos descritivos finais.

| Item | Valve Description |
|---|---|
| 3 | Ball valve, nominal size 2inches, pressure class 300psi, body material in carbon steel ASTM A216 Grade WCB, ball material in stainless steel AISI 316, stem material in stainless steel AISI 316, construction standard as per manufacturer standard, end connections NPT threaded as per ASME B1.20.1. Additional information: API 17292, passage full bore, floating mounted, stem and manual lever operation electrostatically grounded to the body, bidirectional flow, with anti-static device and blow-out proof stem, hydrostatic test as per API598. |
| 4 | Plug valve, nominal size 2inches, pressure class 150psi, body material in carbon steel ASTM A216 Grade WCB, obturator material in stainless steel AISI 316, stem material in stainless steel AISI 316, construction standard as per API599, end connections raised face flange as per ASME B16.5, face-to-face as per ASME B16.10. Additional information: quarter-turn operation with lever, serrated concentric as per MSS SP6, hydrostatic test as per API598 |

Figura 15 - Texto descritivo completo

### 5.2 Comparação com a Extração Manual de Dados

A comparação entre a metodologia automatizada implementada pelo sistema de aprendizado de máquina e o processo manual tradicional destacou diferenças notáveis:

- Eficiência Temporal: Observou-se uma diminuição drástica no tempo necessário para processar propostas. Enquanto o método manual convencional, conforme relatado por profissionais do setor, demandava aproximadamente uma hora e trinta minutos para processar arquivos contendo 10 a 15 itens, a abordagem automatizada realizou a

mesma tarefa em um tempo significativamente menor. Com a automação, a extração dos dados essenciais é concluída em até 30 segundos. Incluindo a revisão subsequente pelo operador e a comparação entre o texto original e o conteúdo extraído, o processo completo requer aproximadamente 5 a 10 minutos, ressaltando uma eficiência temporal notável.

## 6. APLICAÇÃO NO CONTEXTO INDUSTRIAL

### 6.1 Integração com Sistemas de Gerenciamento de Bancos de Dados

A integração do sistema de aprendizado de máquina com sistemas de gerenciamento de bancos de dados é apresentada como fundamental no contexto industrial. Essa sinergia facilita a armazenagem, processamento e recuperação eficiente de informações extraídas das propostas, contribuindo para uma gestão de dados mais eficaz e para a tomada de decisões estratégicas informadas.

### 6.2 Potencial de Escalabilidade e Adaptabilidade

O sistema demonstra um potencial substancial para ampliação e flexibilidade:

- **Escalabilidade Robusta**: Possui a capacidade técnica para processar um volume crescente de dados e acomodar uma gama cada vez maior de tipos de válvulas ou qualquer outro tipo de equipamento sem degradação de desempenho.
- **Adaptabilidade Transversal**: Apresenta a flexibilidade de ser customizado para distintos mercados e idiomas, facilitando a integração global e a penetração em nichos específicos.
- **Flexibilidade Evolutiva**: Está preparado para integrar inovações e avanços tecnológicos em aprendizado de máquina e processamento de linguagem natural (PLN), mantendo-se alinhado com as tendências emergentes e as necessidades do setor.

# 7. CONCLUSÃO

## 7.1 Inovações em Gestão de Propostas Comerciais através de Aprendizado de Máquina

Este estudo destaca-se pela aplicação inovadora de técnicas de aprendizado de máquina na otimização do processo de gestão de propostas comerciais, com um foco específico em Válvulas Industriais e principalmente para o setor de petróleo e gás. Nos domínios caracterizados pela complexidade e a abundância de detalhes técnicos, o sistema proposto emergiu como uma ferramenta valiosa, capaz de processar e analisar informações técnicas com notável precisão e eficiência. A contribuição inovadora deste estudo é reconhecida na sua capacidade de enfrentar e superar desafios específicos destes setores, reformulando significativamente o modo como as corporações abordam e gerenciam a elaboração de propostas comerciais detalhadas e tecnicamente exigentes. Esta abordagem não só aumenta a eficiência e a precisão nas propostas, mas também promove uma gestão mais estratégica e informada nas interações comerciais, estabelecendo um novo padrão no processo de gestão de propostas em ambientes industriais complexos.

## 7.2 Direções para Pesquisa Futura e Extensão de Aplicações

À luz dos resultados alcançados neste estudo, diversas perspectivas emergem para a evolução e aplicação desta pesquisa, particularmente no âmbito das propostas comerciais nas indústrias de óleo e gás e em outros setores industriais. A adoção de modelos avançados de processamento de linguagem, tais como os Generative Pre-train  ed Transformers (GPTs), representa um salto significativo nesse domínio. Estes modelos têm o potencial de transformar a análise de propostas comerciais complexas, desempenhando um papel crucial na interpretação exata de especificações técnicas e requisitos dos clientes.

### 7.2.1 Utilização de LLMs como GPT

O emprego de um LLM (Large Language Model) como o GPT evidencia-se por sua eficácia multifacetada em diversas áreas:

- **Análise e Interpretação de Textos:** LLMs podem ser aplicados para analisar e interpretar extensos volumes de textos, como propostas comerciais, identificando informações-chave e intenções subjacentes.
- **Geração e Aprimoramento de Conteúdo:** Tais modelos auxiliam na geração e no aperfeiçoamento de conteúdo, assegurando precisão técnica e relevância contextual.
- **Assistência na Tomada de Decisão:** O GPT pode ser utilizado para fornecer insights baseados em dados e recomendações, facilitando processos decisórios.

### 7.2.2 Complementaridade com Expressões Regulares

Ao considerar o aprimoramento e enriquecimento de termos em análises complexas, destaca-se a importância da complementaridade entre LLMs e expressões regulares:

- **Enriquecimento de Termos Identificados:** Os LLMs podem enriquecer os termos e conceitos detectados pelas expressões regulares, proporcionando uma análise mais profunda e contextualizada.
- **Ampliação do Escopo de Análise:** A integração de LLMs com expressões regulares permite abranger tanto aspectos estruturais quanto semânticos complexos.

### 7.2.3 Substituição Potencial por LLMs

Em determinados cenários, os LLMs podem substituir expressões regulares, particularmente quando a flexibilidade e o entendimento contextual são essenciais. Entretanto, é crucial avaliar cada situação individualmente:

- **Flexibilidade versus Precisão:** Enquanto LLMs oferecem flexibilidade e compreensão semântica, expressões regulares são mais indicadas para padrões textuais precisos. A decisão dependerá dos requisitos específicos da tarefa.

### 7.2.4 Benefícios Adicionais e Perspectivas Futuras

Este estudo abre perspectivas promissoras para a otimização de propostas comerciais em setores industriais complexos, destacando a integração de tecnologias avançadas como os Generative Pre-trained Transformers (GPTs). Os benefícios e as potenciais aplicações futuras dessa integração são diversos:

- **Minimização de Erros e Aumento da Precisão:** A implementação de GPTs tem como objetivo principal a redução de erros e o aumento da precisão nas propostas. Através da análise detalhada e compreensão aprimorada oferecida por esses modelos, é possível assegurar uma maior exatidão e confiabilidade nas informações apresentadas.
- **Recomendação de Itens Específicos:** Outro aspecto relevante é a capacidade do sistema em sugerir itens específicos, adaptando as propostas conforme as necessidades individuais dos clientes. Esta funcionalidade personalizada não apenas melhora a precisão das propostas, mas também contribui para uma experiência do cliente mais direcionada e satisfatória.
- **Melhoria da Experiência do Cliente:** A elevação na qualidade, velocidade e segurança das propostas, proporcionada pelo uso dos GPTs, tem um impacto direto na experiência do cliente. Ao aprimorar estes aspectos, o sistema busca oferecer um serviço mais eficiente e confiável, resultando em um aumento significativo na satisfação do cliente.

Além desses aspectos, a expansão do escopo de aplicação para diferentes tipos de equipamentos e a integração com sistemas de resposta rápida são vistas como áreas de grande potencial. Avaliações do impacto econômico e ambiental da implementação dessas tecnologias, bem como estudos de longo prazo sobre sua adoção, são componentes essenciais para entender plenamente o alcance e os efeitos dessas inovações.

Em conclusão, a incorporação de GPTs e outras tecnologias avançadas promete transformar não apenas a maneira como as propostas comerciais são criadas e gerenciadas, mas também a forma como as empresas interagem com seus clientes. Esta transformação tem o potencial de elevar significativamente o nível de serviço e satisfação do cliente, posicionando as empresas de forma competitiva no mercado industrial.

### 7.2.5 Limitações do Trabalho

As principais limitações encontradas neste estudo estão ligadas à escassez de grandes conjuntos de dados abertos e acessíveis, um desafio comum no campo de válvulas industriais. Muitas empresas do setor mantêm seus dados confidenciais por razões comerciais e de propriedade industrial, o que restringe significativamente o acesso a informações detalhadas e atualizadas. Essa limitação de dados tem impacto direto na eficácia dos modelos de aprendizado de máquina de grande escala, como os LLMs, que exigem extensos conjuntos de dados para treinamento e aprimoramento de sua precisão e confiabilidade.

Dada a escassez de dados adequados para treinar LLMs, este estudo optou por empregar expressões regulares como ferramenta principal para análise de texto. As expressões regulares, apesar de menos sofisticadas que os LLMs, oferecem uma abordagem mais controlável e direta para a identificação e extração de informações específicas em textos. Elas permitem a implementação de padrões de busca personalizados, adequados para identificar terminologias e especificações técnicas relacionadas a válvulas industriais, mesmo em um conjunto limitado de dados.

Outra limitação observada é a ausência de uma ontologia padronizada e amplamente disseminada para válvulas no mercado consumidor. Isso não apenas gera complicações na troca de informações e dados na indústria, mas também sublinha a necessidade de desenvolver métodos alternativos de coleta de dados. Essas limitações enfatizam a importância de estabelecer parcerias com empresas do setor para acessar dados não públicos sob acordos de confidencialidade, bem como a criação de simulações baseadas em conhecimento técnico. A falta de um esquema de classificação padronizado dificulta a categorização e análise automatizada de informações de válvulas, exigindo sistemas de processamento de dados para serem constantemente verificados e atualizados conforme as inovações do setor. Além disso, a construção de uma ontologia específica para válvulas industriais, com a colaboração de especialistas do setor, representa um passo crucial para padronizar os termos e facilitar a análise automatizada. Enquanto o NIST(National Institute of Standards and Technology), uma agência do Departamento de Comércio dos EUA, trabalha na promoção de padrões e vocabulários legíveis por máquinas, a adoção e disseminação desses padrões no mercado global ainda são desafios a serem superados. Tais esforços são essenciais para reduzir riscos e custos, melhorar

a comunicação e promover inovação e colaboração em um mercado global, sendo fundamentais para o avanço tecnológico e operacional na indústria de válvulas.

## REFERÊNCIAS

**Aggarwal, C. C., & Zhai, C.** (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.

**Aggarwal, C. C., & Zhai, C.** (2012). Mining Text Data. Springer.

**Baeza-Yates, R., & Ribeiro-Neto, B.** (2011). Modern Information Retrieval: The Concepts and Technology behind Search (2nd ed.). Addison-Wesley. ISBN: 978-0131390729

**Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., ... & Trench, M.** (2017). Artificial intelligence: The next digital frontier. McKinsey Global Institute.

**Brynjolfsson, E., & McAfee, A. (2014).** The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company.

**Chowdhury, G. G.** (2003). Natural Language Processing. Annual Review of Information Science and Technology, 37(1), 51-89. https://doi.org/10.1002/aris.1440370103. Acessado em 17 de novembro de 2023.

**Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S.** (2016). Where machines could replace humans—and where they can't (yet). McKinsey Quarterly.

**Cialdini, R.B. (**2006). Influence: The Psychology of Persuasion. Harper Business.

**Christensen, C.M.** (1997). The Innovator's Dilemma. Harvard Business Review Press.

**Cohen, A. M., & Hunter, L.** (2008). Getting started in text mining. PLoS Computational Biology, 4(1), e20. https://doi.org/10.1371/journal.pcbi.0040020. Acessado em 17 de novembro de 2023.

**Fellbaum, C.** (1998). WordNet: An Electronic Lexical Database. MIT Press.

**Goodfellow, I., Bengio, Y., & Courville, A.** (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/. Acessado em 17 de novembro de 2023.

**Hirschberg, J., & Manning, C. D.** (2015). Advances in natural language processing. Science, 349(6245), 261-266. https://doi.org/10.1126/science.aaa8685. Acessado em 17 de novembro de 2023.

**Jurafsky, D., & Martin, J. H.** (2019). Speech and Language Processing (3rd ed.). Prentice Hall. ISBN: 978-0131873216

**Kotler, P., & Keller, K.L.** (2016). Marketing Management. Pearson.

**Kuckartz, Udo, e Rädiker, Stefan.** "Coding Text and PDF Files". Springer International Publishing, 2019.

**Liddy, E. D.** (2001). Natural Language Processing. In B. Cronin (Ed.), Encyclopedia of Library and Information Science (2nd ed., Vol. 2, pp. 2046-2056). Marcel Dekker. ISBN: 978-0824720797

**Link Apress.** "Processing Text Files: Text Is Everywhere". 2009.

**Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., & Marrs, A.** (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. McKinsey Global Institute.

**Manning, C. D., Raghavan, P., & Schütze, H.** (2008). Introduction to Information Retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/. Acessado em 17 de novembro de 2023.

**Muse, Amiin.** "AD Elog Data Search Using Natural Language Processing Techniques". US DOE, 2023.

**Navigli, R., & Ponzetto, S. P.** (2012). BabelNet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, 217-250.

**Palmer, David D.** "Text Preprocessing". Chapman and Hall/CRC, 2020.

**Porter, M. E., & Heppelmann, J. E.** (2014). How smart, connected products are transforming competition. Harvard Business Review, 92(11), 64-88.

**Puri, S., & Singh, S. P.** (2016). A technical study and analysis of text classification techniques in N - Lingual documents. IEEE.

**Ries, E.** (2011). The Lean Startup. Crown Publishing Group.

**Rizvi, S. T. R., Khan, M. A., & Khan, M. S.** (2018, March). Ontology-based information extraction from technical documents. In 2018 13th International Conference on Advanced Communication Technology (ICAART) (pp. 493-500). IEEE.

**Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., & Harnisch, M.** (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. Boston Consulting Group, 9(1), 54-89.

**Schwab, K.** (2016). The fourth industrial revolution. World Economic Forum.

**Sebastiani, F.** (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1-47. https://doi.org/10.1145/505282.505283. Acessado em 17 de novembro de 2023.

**Singh, Jyotika.** "Data Preprocessing and Transformation". Chapman and Hall/CRC, 2023.

**Stein, Aviel J.** "Applying Natural Language Processing Techniques to Code". Drexel University Libraries, 2022.

**Tang, B., Feng, Y., & Wang, X.** (2008). A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. Journal of Cheminformatics, 10(1), 1-10.

**Tang, J., Zhang, D., & Yao, L.** (2008). Social Network Extraction of Academic Researchers. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08).

**Westerman, G., Calméjane, C., Bonnet, D., Ferraris, P., & McAfee, A.** (2014). Digital transformation: A roadmap for billion-dollar organizations. MIT Center for Digital Business and Capgemini Consulting.

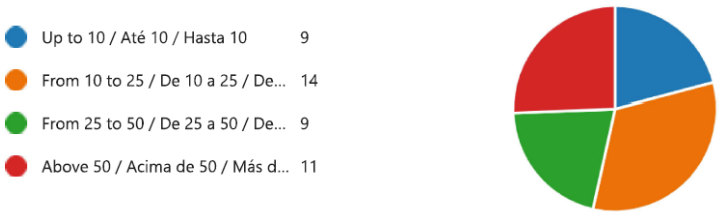
## APÊNDICE A – PESQUISA REALIZADA COM FABRICANTES E DISTRIBUIDORES DE VÁLVULAS

# Survey on Communication and Data Processing in Industrial Valve Quotations

| 43 | 09:51 | Active |
|---|---|---|
| Responses | Average time to complete | Status |

1. Is your company a manufacturer of industrial valves, or are you solely a distributor? A sua empresa é fabricante de válvulas industriais ou apenas distribuidora? ¿Su empresa es fabricante de válvulas industriales o solo distribuidora?


2. How many quotations do you individually receive weekly on average? / Quantas cotações você recebe semanalmente em média individualmente? / ¿Cuántas cotizaciones recibe semanalmente en promedio individualmente?




3. What is the average number of items/descriptions per quotation? / Qual é a quantidade média de itens/descritivos por cotação? / ¿Cuál es la cantidad promedio de ítems/descriptivos por cotización?

4. How long does it take you to prepare a quote for 10 items that are already known, from the moment of receipt until dispatch? / Quanto tempo você leva para preparar uma cotação de 10 itens já conhecidos, do momento do recebimento até o envio? / ¿Cuánto tiempo te lleva preparar una cotización de 10 artículos ya conocidos, desde el momento de la recepción hasta el envío?

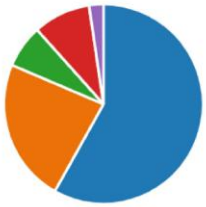
| | |
|---|---|
| Up to 1 hour / Até 1 hora / Hast... | 26 |
| From 1 to 3 hours / De 1 a 3 hor... | 9 |
| From 3 to 5 hours / De 3 a 5 hor... | 4 |
| More than 5 hours / Mais de 5 h... | 4 |

5. In which format do you receive the quotations? / Em qual formato você recebe as cotações? / ¿En qué formato recibe las cotizaciones?

| | |
|---|---|
| PDF | 29 |
| XLSX (EXCEL) | 28 |
| DOCX (WORD) | 9 |
| HTML(in the body of the email/... | 14 |
| Other | 5 |

6. Which file extension requires more effort to process? / Qual extensão de arquivo requer mais esforço para processar? / ¿Qué extensión de archivo requiere más esfuerzo para procesar?

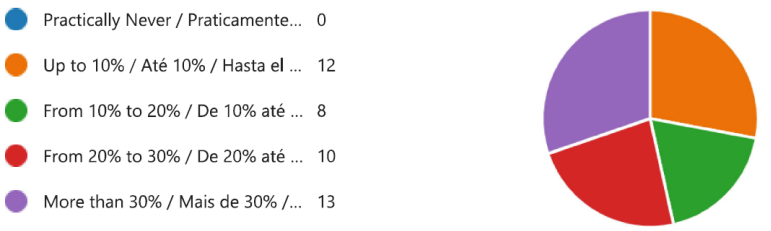
| | |
|---|---|
| PDF | 25 |
| XLSX (EXCEL) | 10 |
| DOCX (WORD) | 3 |
| HTML(in the body of the email/... | 4 |
| Other | 1 |



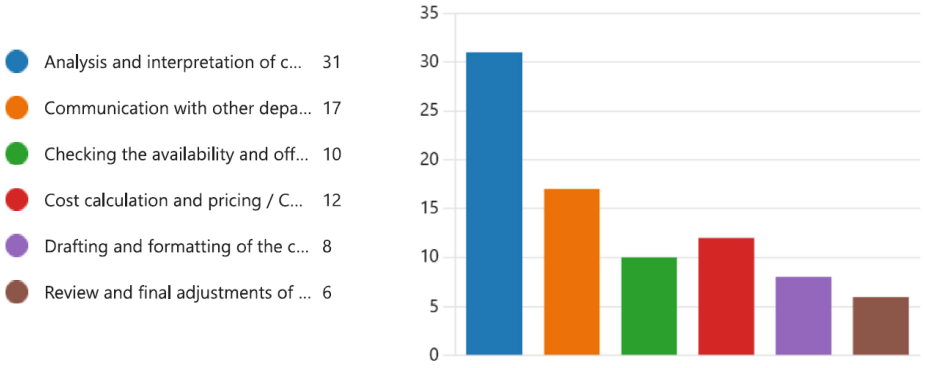
7. Regarding the descriptive texts of the valves, what makes your work more difficult? / Em relação aos textos descritivos das válvulas, o que mais dificulta o seu trabalho? / En cuanto a los textos descriptivos de las válvulas, ¿qué dificulta más su trabajo?

| | |
|---|---|
| Illegible abbreviations / Abrevia... | 13 |
| Descriptive in another language... | 13 |
| Descriptions that are too simplif... | 32 |
| Other issues / Outros problema... | 3 |

8. How often do you receive incomplete specifications in the list of quoted items? / Com que frequência você recebe especificações incompletas na lista de itens cotados? / ¿Con qué frecuencia recibe especificaciones incompletas en la lista de artículos cotizados?




9. In which of the following stages of the quotation process do you generally invest the most time? / Em qual das seguintes etapas do processo de cotação você geralmente investe mais tempo? / ¿En cuál de las siguientes etapas del proceso de cotización generalmente invierte más tiempo?



## APÊNDICE B – DETALHES DE IMPLEMENTAÇÃO DO CÓDIGO

```
# Importações
import re
import nltk
import pandas as pd
import csv
from io import StringIO
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from collections import defaultdict

# Baixar o pacote NLTK necessário
nltk.download('punkt')

# Função para ler um arquivo PDF e retornar seu texto
def ler_pdf(caminho_arquivo):
    output_string = StringIO()
    with open(caminho_arquivo, 'rb') as f:
        parser = PDFParser(f)
        doc = PDFDocument(parser)
        rsrcmgr = PDFResourceManager()
        dispositivo = TextConverter(rsrcmgr, output_string, laparams=LAParams())
        interpretador = PDFPageInterpreter(rsrcmgr, dispositivo)
        for pagina in PDFPage.create_pages(doc):
            interpretador.process_page(pagina)
    return output_string.getvalue()

# Solicitar entrada do usuário para escolher o tipo de arquivo
escolha = input("Escolha a fonte de dados (Digite 'excel', 'txt' ou 'pdf'): ").strip().lower()
texto = ""

# Processamento condicional com base na escolha do usuário
if escolha == 'excel':
    arquivo_excel = "caminho_para_arquivo_excel.xlsx"
    df = pd.read_excel(arquivo_excel)
    texto = ' '.join(map(str, df[df.columns[0]].tolist())).lower()

elif escolha == 'txt':
    try:
        with open("caminho_para_arquivo_texto.txt", "r", encoding="utf-16") as f:
```

```
            texto = f.read().lower()
    except UnicodeDecodeError:
        print("Erro ao ler o arquivo com codificação utf-16. Por favor, verifique a codificação do arquivo.")
        exit()

elif escolha == 'pdf':
    arquivo_pdf = "caminho_para_arquivo_pdf.pdf"
    texto = ler_pdf(arquivo_pdf).lower()

else:
    print("Escolha inválida. Por favor, digite 'excel', 'txt' ou 'pdf'.")
    exit()

# Definir o padrão para corresponder às descrições das válvulas
padrao = r"(válvula gaveta.*?|válvula esfera.*?|válvula globo.*?|válvula retenção.*?|válvula macho.*?)(?= válvula gaveta| válvula esfera| válvula globo| válvula retenção| válvula macho|$)"

frases_valvulas = re.findall(padrao, texto, re.DOTALL)
frases_filtradas = [frase for frase in frases_valvulas if len(frase.split()) > 3]
print(" <> ".join([frase.strip() for frase in frases_filtradas]))

# Processamento adicional do texto
tipos_valvulas = ["elevação", "válvula gaveta", "válvula esfera flutuante", "válvula esfera trunnion", "válvula esfera", "válvula globo", "válvula retenção oscilante", "válvula retenção pistão", "válvula retenção dupla placa", "válvula retenção", "válvula macho"]
for valvula in tipos_valvulas:
    texto = re.sub(valvula, " " + valvula, texto)

criterio_divisao = r"(?= elevação| válvula gaveta| válvula esfera flutuante| válvula esfera trunnion| válvula esfera| válvula globo| válvula retenção oscilante| válvula retenção pistão| válvula retenção dupla placa| válvula retenção| válvula macho)"
frases = re.split(criterio_divisao, texto)
frases_filtradas = [frase for frase in frases if len(frase.split()) > 3]

# Tokenização e lematização das frases
lemmatizer = WordNetLemmatizer()
frases_tokenizadas = [word_tokenize(frase) for frase in frases_filtradas]
frases_lematizadas = [" ".join([lemmatizer.lemmatize(token) for token in tokens]) for tokens in frases_tokenizadas]

# Filtrar frases contendo a palavra "válvula" e suas derivações
frases_valvulas = [frase for frase in frases_lematizadas if re.search(r'\bválvula\b', frase)]

# Vocabulário
vocabulary = {
```

"valve_type": ["ball valve", "gate valve", "globe valve", "check valve", "plug valve", "butterfly valve", "control valve", "needle valve", "relief valve", "solenoid valve", "strainer"],
"nominal_size": ["ø 1/4", "ø 3/8", "ø 1/2", "ø 3/4", "ø 1", "ø 1 1/4", "ø 1 1/2", "ø 2", "ø 2 1/2", "ø 3 ", "ø 4", "ø 5", "ø 6", "ø 8", "ø 10", "ø 12", "ø 14","ø 16", "ø 18", "ø 20", "ø 22", "ø 24", "ø 26", "ø 28", "ø 30", "ø 32", "ø 34", "ø 36", "ø 38", "ø 40", "ø 42", "ø 44", "ø 46", "ø 48", "ø 50", "ø 52", "ø 54", "ø 56", "ø 58", "ø 60", "ø 62", "ø 64", "ø 66", "ø 68", "ø 70", "ø 72", "ø 74", "ø 76", "ø 78", "ø 80", "ø 82", "ø 84", "ø 86", "ø 88", "ø 90"],
"pressure_class": ["class 125#", "class 125psi", "class 125", "125# class", "class 125 psi", "125psi", "125#", "125 psi", "125# class", "cl125", "cl.125", "cl. 125", "cl 125psi", "cl 125#", "cl. 125#", "class 150#", "class 150psi", "class 150", "150# class", "class 150 psi", "150psi", "150#", "150 psi", "150# class", "cl150", "cl.150", "cl. 150", "cl 150psi", "cl 150#", "cl. 150#", "class 300#", "class 300psi", "class 300", "300# class", "class 300 psi", "300psi", "300#", "300 psi", "300# class", "cl300", "cl.300", "cl. 300", "cl 300psi", "cl 300#", "cl. 300#", "class 600#", "class 600psi", "class 600", "600# class", "class 600 psi", "600psi", "600#", "600 psi", "600# class", "cl600", "cl.600", "cl. 600", "cl 600psi", "cl 600#", "cl. 600#", "class 800#", "class 800psi", "class 800", "800# class", "class 800 psi", "800psi", "800#", "800 psi", "800# class", "cl800", "cl.800", "cl. 800", "cl 800psi", "cl 800#", "cl. 800#", "class 900#", "class 900psi", "class 900", "900# class", "class 900 psi", "900psi", "900#", "900 psi", "900# class", "cl900", "cl.900", "cl. 900", "cl 900psi", "cl 900#", "cl. 900#", "class 1500#", "class 1500psi", "class 1500", "1500# class", "class 1500 psi", "1500psi", "1500#", "1500 psi", "1500# class", "cl1500", "cl.1500", "cl. 1500", "cl 1500psi", "cl 1500#", "cl. 1500#", "class 2500#", "class 2500psi", "class 2500", "2500# class", "class 2500 psi", "2500psi", "2500#", "2500 psi", "2500# class", "cl2500", "cl.2500", "cl. 2500", "cl 2500psi", "cl 2500#", "cl. 2500#", "3000", "3000psi", "3000#", "class 3000", "5000", "5000psi", "5000#", "class 5000", "10000", "10000psi", "10000#", "class 10000", "15000", "15000psi", "15000#", "class 15000", "20000", "20000psi", "20000#", "class 20000"],
"body_material": "f304", "f304l", "f304h", "f316", "f316l", "f316h","f321", "f347", "f347h", "f317", "f317l", "f310", "f321h", "f44", "f51", "f53", "f55", "f5", "f9", "f11", "f22", "f91", "cf8", "cf8m", "cf8c", "cf3", "cf3m", "cg8m", "cn7m", "ck3mcun", "cd4mcu", "ce8mn", "cd3mn", "lcb", "lcc", "lc2", "lc3", "ca6nm", "ca15", "c5", "c12", "c12a", "wcb", "wcc", "wc6", "wc9", "c5", "c12", "c12a", "cd3mwcun", "ce3mn", "cd6mn", "cd3mn", "cd4mcun", "cb7cu-1", "cb7cu-2", "cf10smnn"],
"end_connections": ["raised face", "rf", "socket weld", "socket welding ends", "sw", "rtj", "ring joint type", "bw", "butt weld", "butt welding ends", "nipple", "thread npt", "npt threaded", "npt", "male", "female"],
"other_terms": ["body", "cover", "ball", "stem", "seat", "ring", "bolt", "flange", "gasket", "packing", "backseat", "bonnet", "trim", "actuator", "disc","handwheel", "yoke", "spindle", "bushing", "gland", "valve plate", "orifice", "port", "butterfly", "diaphragm", "plug", "sleeve", "lever", "gear operator", "position indicator", "pressure seal", "bellows", "flow coefficient", "cavitation", "choke", "trim material", "flow direction", "pressure rating", "end connection", "leakage class", "ISO standard", "API specification", "fugitive emissions", "torque", "bypass valve", "valve body material", "wedge", "butterfly valve", "gate valve", "globe valve", "ball valve", "plug valve", "pneumatic actuator", "hydraulic actuator", "electric actuator", "fire safe design", "cryogenic service", "corrosion resistance", "pressure drop", "valve sizing", "stem packing", "seat leakage", "API 6D", "API 600", "API 602", "API 607", "API 622", "valve testing", "hydrostatic testing", "pneumatic testing", "actuator sizing", "face to face dimensions", "flange rating", "bore size", "trim number", "shutoff class", "pipe

```
size", "valve orientation", "flow rate", "temperature rating", "design standard", "maintenance procedure", "installation guide", "failure mode", "wear resistance", "hard facing"]

}

# Coletar palavras do vocabulário de cada frase
palavras_vocabulario = defaultdict(list)
for frase in frases_valvulas:
    for categoria, termos in vocabulario.items():
        for termo in termos:
            if termo in frase:
                palavras_vocabulario[frase].append(termo)

# Contar tipos de válvulas
contagem_tipos_valvulas = defaultdict(int)
for frase, termos in palavras_vocabulario.items():
    for termo in termos:
        if termo in vocabulario["valve_type"]:
            contagem_tipos_valvulas[termo] += 1

# Imprimir resultados
for idx, (frase, termos) in enumerate(palavras_vocabulario.items(), 1):
    print(f"{idx}. {frase} -> {termos}")

print("\nContagem de Tipos de Válvulas:")
for tipo_valvula, contagem in contagem_tipos_valvulas.items():
    print(f"{tipo_valvula}: {contagem}")
```