UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Machine Learning: Abordagem de identificação de oportunidades de negócios em portais B2B

Luiz Fernando Duque Estrada Antelo

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Machine Learning: Abordagem de identificação de oportunidades de negócios em portais B2B

USP - São Carlos

2023

Luiz Fernando Duque Estrada Antelo

Machine Learning: Abordagem de identificação de oportunidades de negócios em portais B2B

Trabalho de conclusão de curso apresentado 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. Fernando Osório

USP - 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)

D946m

Duque Estrada Antelo, Luiz Fernando
Machine Learning: Abordagem de identificação de oportunidades de negócios em portais B2B / Luiz Fernando Duque Estrada Antelo; orientador Prof Dr Fernando Osório. -- São Carlos, 2023.
49 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. INTELIGÊNCIA ARTIFICIAL. 2. APRENDIZADO COMPUTACIONAL. 3. NEGÓCIOS. I. Osório, Prof Dr Fernando, 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

DEDICATÓRIA

A meus pais e minha esposa pelo apoio e presença em todos os momentos juntos vividos, fundamentais para chegar até aqui.

AGRADECIMENTOS

Aos meus pais, que nunca mediram esforços, para que meus irmãos e eu tivéssemos acesso à educação que buscávamos, se esforçaram e me deram meu primeiro computador ainda quando criança, e sem ter ideia como isso foi poderoso e que me permitiu trilhar meu caminho, abrindo assim, um universo de possibilidades até os dias de hoje e concluir esta monografia.

À minha esposa, parceira e companheira de toda uma vida, que está sempre a meu lado em todos os momentos felizes e perrengues.

À professora Dra. Solange Rezende, que com empatia recebeu todos os desafios conspirados pelo universo ao longo deste curso e transformou em palavras assertivas de incentivo, sempre me incentivando a olhar para frente.

Ao Professor Dr. Fernando Osório que conduziu nossas discussões de forma leve e construtiva, com conselhos práticos e objetivos, me trouxe calma e confiança diante das incertezas. A quem ainda sou devedor de um *capuccino* pelas trocas bem-humoradas.

À Professora Dra. Roseli Romero que nos orientou e nos conduziu de volta ao curso em vários momentos da nossa jornada no MBA.

EPÍGRAFE

Life takes on a different aspect when you step out with decision and purpose.

Ralph C. Smedley

RESUMO

ANTELO, L. F. D. E. **Machine Learning: Abordagem de identificação de oportunidades de negócios em portais B2B**. 2023. 49 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

O aumento das relações comerciais intermediada por plataformas eletrônicas entre empresas (B2B), gera de forma crescente cada vez mais volumes de dados em forma de detalhamentos e descrições de requisições de compra de bens e serviços. Mais comumente, empresas adotam os processos B2B em busca da racionalização de processos e margens operacionais, publicando suas demandas de aquisições, a fim de atingir o maior número possível de fornecedores gabaritados. Essa modalidade de compra oferece enorme oportunidade para todas as empresas dispostas a coletar, analisar e apresentar propostas comerciais em tempo hábil para atender a demanda de seus clientes. O presente trabalho estudou e implementou uma forma de automatizar a coleta de ofertas de negócio publicadas em portais eletrônicos de licitações eletrônicas. Utilizando um *web-crawler* para coleta das demandas de compra e aprendizado de máquina para classificação da potencial área de solução demandada através de um classificador multinomial Naive Bayes para classificá-las de forma mais eficiente, acelerando o envio para as equipes comerciais competentes, oferecendo mais tempo e chance de qualificar as oportunidades e produzir melhores propostas para a conversão em negócios o que resultou no aumento de 75% volume médio semanal de oportunidades disponíveis as equipes e 3 potenciais oportunidades de negócio durante o período de 10 semanas do estudo.

Palavras-chave: Inteligência Artificial; Aprendizado de Máquina; linguagem python;

Compras Eletrônicas; B2B; classificador multinomial Naive Bayes;

ABSTRACT

ANTELO, L. F. D. E. **Machine Learning: Business Opportunity Identification Approach on B2B Portals**. 2023. 49 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The rise of commercial relationships over electronic procurement platforms among enterprises, generates a growing volume of requests for proposals and details of goods and services. Commonly, enterprises adopt B2B processes in search of streamlined processes and better operational margins, publishing their acquisition demands, to attain the largest number possible of eligible suppliers. This way of doing business offers huge opportunities for all enterprises willing to collect, analyze and present proposals in a timely manner to meet clients' demands. The present work studied and implemented an automated data scraping using a web-crawler to collect procurement demands and machine learning to classify potential areas of business solution using a multinomial Naive Bayes classifier for a more efficient identification, accelerating its routing to commercial teams, offering them more time to develop a better proposal, thus better chance to convert it into business, yielding in an average growth of 75% of opportunities available for sales teams on a weekly volume and 3 potential opportunities being considered during the 10 week period the study took place.

Keywords: Artificial Intelligence; Machine Learning; python language; B2B; multinomial Naive Bayes classifier.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ing.	–	tradução do inglês
eProcurement	–	ing. processo eletrônico de compras.
B2B	–	*Business to Business* (ing. de empresas para empresas)
B2C	–	*Business to Consumer* (ing. de empresas para consumidores)
ePV	–	eProcurement Vendor (ing. empresa especializada no processo de transações eletrônicas)
API	–	*Application Programming Interface – tecnologia de interconexão entre aplicações;*
ESG	–	ing. práticas de governança corporativa (*Enterprise Social Governance*)
ERP	–	ing. sistemas de gestão de recursos corporativos (*Enterprise Resource Planning*)
IT	–	Information Technology (ing. Tecnologia da Informação ou TI)
AI	–	*Artificial Intelligence* (ing. Inteligência Artificial)
Keyword	–	ing. Palavra-chave que se relaciona a um tópico ou assunto
XLSX	–	Formato tabular padrão do MS Excel

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização do Problema

Já há algum tempo o processo de compra das empresas e seu relacionamento com fornecedores vem convergindo para plataformas eletrônicas de compras e segundo estudo do IDC, “75% dos compradores B2B entre empresas buscam novos fornecedores em plataformas sociais” (Shanks, 2016, p14). Esta mudança ocorre devido à necessidade de atender com maior agilidade as demandas de compras, integração aos processos e controle de gastos, além de fatores sociológicos, intensificados após a pandemia de COVID-19, como exemplos: “[...] o eCommerce no Brasil cresceu 41% em vendas (R$ bi) em 2020 e o país ficou em 4º lugar com 27% de crescimento de comércio eletrônico mundialmente, [...] no topo desse crescimento *Marketplaces* cresceram ainda mais 52%. Com esse patamar, o formato foi responsável por 84% das vendas online. Fonte: Ebit | *Nielsen Webshoppers* 43 Inteligência de Mercado | Globo” (Simões; Jorquera, 2021)

Com as mudanças comportamentais, as experiências pessoais em processos eletrônicos vêm se fundindo com a experiência corporativa. Hábitos pessoais, como por exemplo: buscar referências de produtos em redes de compras e páginas de fabricantes, assinar documentos eletronicamente, obter documentos em via somente eletrônica, notificação eletrônica de recebimento de bens e serviços, usar um certificado digital para oficializar um documento etc. vem sendo absorvida pelas empresas como reflexo do que aceitamos como prática corrente nas transações (Simões; Jorquera, 2021).

Recentemente as empresas, por crescente pressão em busca de margens operacionais e competitividade, buscam a racionalização de custos através da implementação de plataformas de gestão corporativa (ERP). Estas plataformas, em geral, são modulares e podem permear desde a gestão de produção e do fluxo financeiro, até o processo eletrônico de compras ou ing. *eProcurement* (Mercado Eletrônico, 2023), promovendo a integração e a visão executiva de gestão, do ponto de vista de receitas, gastos, recebíveis etc.

Durante o período da pandemia do COVID-19, notamos um aumento prático na adoção de soluções de ERP’s entre os clientes atendidos, ou pelo menos a implementação do módulo de compras. Este aumento também foi acarretado pela possibilidade, cada vez mais crescente, de interconexão entre as empresas, tanto por plataformas específicas, quanto interfaces programáveis (API), pois permitem que diversos sistemas se comuniquem com facilidade. A

adoção de ERP's tem sido um motivador para a implementação de plataformas de B2B pelas empresas ou até mesmo a migração para plataformas integradas/marketplaces.

Tradicionalmente, existem duas formas de implementar o processo de *eProcurement*:

1) Adoção de um provedor de serviço de compras especializado (*eProcurement Vendor ou ePV*): que faz o papel de conectar empresas e provedores/fornecedores. Com o passar do tempo, cada vez mais, está se convertendo em plataformas de conexão de múltiplos fornecedores, como um *marketplace* (portal para múltiplas empresas transacionarem), como por exemplo o Mercado Eletrônico (https://www.me.com.br)

2) Por plataforma dedicada: desenvolvida pelas empresas que no passado dispunham de equipes dedicadas a manter esses ambientes como parte do seu custo de IT. Muitas vezes, quando as empresas implementam o ERP, dão lugar ao módulo de compras desse tipo de solução. Algumas empresas cresceram tanto suas equipes, que, em algum momento, precisaram desmembrá-las como fornecedores de mercado (ePV). Este movimento no mercado é chamado de *"spin-off"*, converte esses serviços em clientes únicos ou compartilhados, onde a estratégia é dividir ou levar aos fornecedores o custo desta operação de compra eletrônica.

Com essa evolução no processo de compra das empresas, as áreas comerciais perceberam que existe uma oportunidade em se relacionar com clientes de forma mais eletrônica, pois, para determinados tipos de compras de produtos e serviços, pode-se prestar seu papel de forma eficiente. Essa forma de relacionamento comercial eletrônico ou busca por potenciais fornecedores em plataformas sociais é denominada *Social Selling* ou *Digital Selling* (Shanks, 2016).

Neste modelo, as oportunidades podem ser identificadas, e o vendedor pode educar ou conectar-se com os influenciadores/tomadores de decisão dentro das empresas via plataformas como *Linkedin*, para determinado segmento ou até identificar empresas que possuem seus processos de B2B para tornar-se um fornecedor gabaritado e monitorar as demandas de compras que surjam na plataforma.

A migração para plataformas eletrônicas torna esse processo interessante, pois um crescente volume de oportunidades fica disponível para quem estiver interessado e disposto em capturá-las cadastrando sua empresa nos portais de B2B e aderindo ao seu processo de validação. Por outro lado, existe o desafio de uma vez identificadas as oportunidades, estas

cheguem para as equipes internas, para que trabalhem as propostas e retornem ao cliente em tempo hábil. Quanto mais rápida for feita a identificação e classificação das oportunidades, mais tempo pode ser dedicado para o trabalho estratégico, para se montar e apresentar uma proposta de valor ao cliente, com melhores chances de conversão dela em negócio.

1.2 Justificativa e motivação

Muitas oportunidades são publicadas no formato de pregão eletrônico ou licitação, ou seja, podendo ser um leilão em tempo real ou processo de compra/edital, onde estão disponíveis na maioria das vezes* sem reserva para qualquer empresa. Porém o volume de processos é cada vez mais alto e esbarra na necessidade de se validar de forma assertiva se o escopo da oferta é válido ou está fora do foco de atuação da empresa.

O menor tempo decorrido para se identificar, classificar e mobilizar as equipes necessárias, ou para redirecionar à uma empresa parceira de negócio que seja especializada em determinada tecnologia, pode aumentar muito a chance de conversão de negócios.

Figura 1 Etapas desde a identificação de oportunidade até apresentação de proposta

Fonte: o autor

Portanto, a motivação do presente estudo é automatizar o processo de captura das oportunidades e usar uma técnica de AI para ganhar escala e automatizar o processo de coleta, preparação, identificação e classificação de oportunidades selecionadas das plataformas de B2B, para levar estas oportunidades em menor tempo possível à equipe comercial, para que estas possam qualificar e progredir para converter a oportunidade em um negócio ou direcionar a oportunidade a um parceiro comercial.

- Em alguns processos do governo é feita uma priorização a microempresas (ME) como forma de incentivar o fornecedor local, que é identificado na licitação.

1.3 Resultados e impactos esperados

Com este estudo esperou-se estabelecer uma abordagem prática, para coletar de forma automatizada, depurar, classificar e rotear potenciais oportunidades de licitação de compras para atuação das equipes comerciais em um tempo mínimo, permitindo que eles tenham o máximo de tempo possível para trabalhar as oportunidades identificadas e evoluírem as propostas com maior potencial de sucesso.

Como indicadores para quantificar o projeto foram selecionados: volume semanal de licitações coletadas, o número de oportunidades identificadas por profissional e oportunidades identificadas.

Buscou-se assim, o aumento do volume inicial de oportunidades identificadas com ganhos de escala, para alimentar o funil de vendas, e, com isso, espera-se aumentar a taxa de identificação, conversão e conclusão de negócios interessantes para a empresa.

1.4 Organização do trabalho

O trabalho está organizado em 4 capítulos. No capítulo 1 apresenta-se todos os aspectos introdutórios. No capítulo 2, as fundamentações de embasamento dos temas, tanto teórico quanto dos temas de negócios abordados. No capítulo 3 é feito o desenvolvimento prático da identificação do problema, tomada de decisões de implementação e os resultados obtidos.
No capítulo 4 são apresentadas conclusões gerais e pontos de interesse em pesquisa futura a partir do que foi encontrado.

2 Fundamentação Teórica

2.1 Fundamentos relacionados ao tema

Hoje em dia o volume de informação disponível na internet em qualquer campo de atuação demanda um modelo automatizado para varredura e coleta comumente chamado de *Data-Crawler* (data-*scraper* ou de forma curta, *scraper* ou *crawler*) que é um motor de navegação nas páginas de conteúdo usando uma lógica especificada, varrendo os elementos presentes nas páginas (campos, tabelas, documentos etc.) em busca de dados pertinentes. Com isso objetiva-se economizar uma parcela significativa de trabalho manual dos dados para tratamento e análise.

Um exemplo de crawler é o Playwright (https://playwright.dev/) que inicialmente foi desenvolvido para ser usado como um automatizador de testes em interfaces web, que possui grande compatibilidade com os principais navegadores e sistemas operacionais, incluindo suas versões mobile e aceita scripts de teste e que ao final pode capturar dados dos elementos das páginas, oferecendo assim a capacidade de coletar e armazenar essas informações.

O Processamento de Linguagem Natural (NLP) comumente é feito a partir da vetorização dos textos, para se obter o conjunto mais determinante. Removendo-se as palavras de maior frequência (*stopwords)* através da tokenização das palavras relevantes usando-se um modelo simples como o de Zipf.

O algoritmo de Naive Bayes é amplamente utilizado em aplicações de classificação discreta e se baseia em uma interpretação do teorema de Bayes e na presunção da hipótese que os atributos de uma classe de elementos são condicionalmente independentes. O que na prática

muitas vezes não acontece, mas que entrega boa acurácia e eficiência computacional (Webb, 2016).

O teorema de Bayes na sua forma direta estabelece a probabilidade de um atributo y dado x:

$$\mathrm{P}(y \mid \mathbf{x}) = \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y)/\mathrm{P}(\mathbf{x})$$

Usando-se a hipótese de que os atributos são condicionalmente independentes dada uma classe, para o dado atributo que se quer buscar ela pode ser reescrita em forma do produtório:

$$\mathrm{P}(\mathbf{x} \mid y) = \prod_{i=1}^{n} \mathrm{P}(x_i \mid y)$$

Onde x_i é o valor do enésimo atributo de x e n é o número total de atributos, simplificando-se:

$$\mathrm{P}(\mathbf{x}) = \prod_{i=1}^{k} \mathrm{P}(c_i)\mathrm{P}(\mathbf{x} \mid c_i)$$

Onde k é o número de classes e c_i é a enésima classe. A equação inicial do teorema então pode ser calculada normalizando-se os numeradores do lado direito. Essa simplificação torna o classificador um modelo linear que para atributos categóricos, as probabilidade P(y) e $P(x_i \mid y)$ normalmente são derivadas das frequências de contagens armazenados em vertores, varrendo-se o texto no momento de treinamento do modelo. À medida que o modelo vai iterando esse vetor de frequência poder ser atualizado dado seu custo computacional linear sem perda de muita performance, suportando um aprendizado incremental.

O modelo multinomial de Naive Bayes considera a informação do número de vezes uma palavra (ou *token*) aparece no texto considerado. Ele trata cada ocorrência do token como um evento separado que são assumidos como independentes entre eles.

O aprendizado de máquina neste caso oferece uma implementação adequada para o objeto de texto a ser analisado, que possui universo bem definido além de poder ser incrementado a cada iteração semanal.

3 Metodologia e Desenvolvimento

3.1 Identificação do problema

Muitos portais B2B oferecem a publicação de oportunidades de negócio, porém os volumes de informação tornam o processo de captura, tratamento e análise das oportunidades com base em seus enunciados e componentes uma tarefa inviável em um prazo exíguo que se necessita para poder engajar as equipes de trabalho, formular uma proposta comercial e retornar a empresa demandante dentro do prazo estipulado por ela.

A automação de todo o processo inicial até o envio às equipes internas, constitui diferencial competitivo que permite mais tempo sendo usado na parcela mais nobre do trabalho de avaliação e elaboração de proposta dentro do prazo necessário, aumentando as chances de sucesso em desenvolver novos negócios.

O problema foi dividido de forma macro em A) coleta das informações para as bases de dados, B) Tratamento dos dados e C) Classificação e tomada de decisão.

Como uma das propostas iniciais do projeto era sempre abordar o processo de forma prática e incremental, adotamos uma abordagem ágil com ciclos semanais (*sprints)* onde o objetivo era agregar valor a cada ciclo e que possibilitasse sempre ganho de escala. Desta forma testando cada hipótese com menor esforço possível para então decidir pela implementação e próximo passo.

3.2 Base de dados utilizada no experimento

Este projeto utilizou o portal Compras.gov.br do governo brasileiro por se tratar de portal de publicação aberto, devido à lei de transparência e por abarcar muitas esferas e institutos governamentais, buscando-se ganho da curva de aprendizado com a eventual implementação inicial (Rauen, A. T. O. 2022).

Figura 2 Portal de compras do governo federal

Fonte: https://www.gov.br/compras/pt-br

Ao iniciar a exploração do portal escolhido, foi levantada das informações disponíveis de identificação de licitação, tais como: Órgão comprador, código UASG da instituição compradora, Número do Edital, Objeto da licitação, Data da abertura, Data limite para entrega, Endereço completo e link do edital. A coleta de dados e a tabulação inicial foi feita de forma manual, o que constituiu o início da base de dados utilizada. À medida que o processo foi automatizado, a cada semana de interação, a base foi agregada de forma incremental.

Para se identificar as possíveis oportunidades através das buscas manuais foi criado um arquivo de definição de palavras-chave (*Keywords)* que remetiam os temas. Essa lista foi concebida por processo coletivo de ideias *(brainstorming)*, bem como coleta com as equipes de cada áreas solicitando termos específicos relacionados à natureza dos produtos e consolidadas em uma tabela (ver apêndice A)

Tabela 2 Amostra *keywords* de busca

Sustaintability	Security	Cloud	Automation	DATA & AI	Turbo/Instana	Storage
Gestão de ativos	Endpoint Detection and Response (EDR)	PaaS	AIOps	observabilidade	Armazenamento	Renew
Sustentabilidade	controle de acesso	SaaS	fluxo de trabalho	desempenho	Backup	Revenue
pegada de carbono	MaaS 360	IaaS	workflow	data lake	automação	DataCenter
cadeia de suprimento	GDPR	Nuvem	RPA	Ética	application Resource Management (ARM)	Backlevel
supply chain	privacidade	hybrid	Integração de Dados	Qualidade de dados	gestão de recursos	Kubernetes

Fonte: o autor

Para melhor parametrizar as oportunidades, os seguintes campos foram adicionados aos dados coletados: Data da pesquisa, *Keyword* original de busca, Área de negócio associada a esta palavra-chave;

3.3 Pré-processamento

A cada iteração semanal, muitas licitações eram consideradas e precisavam ser tratadas devido a casos de:

1) **Duplicidade:** uma licitação era capturada por buscas distintas pois possuíam diferentes keywords, ex: "Compra de licenças de software de *Endpoint Security Management*" era capturada pela busca de ***Security*** e por outra que continha ***Management.*** Nestes casos uma de-duplicação foi aplicada usando-se o método *drop_duplicates* do Pandas usando o objeto descritivo da licitação;
2) **Falsos positivos:** como um exemplo ilustrativo buscas pela *keyword* "Servidor" resultaram em muitas licitações em que esta palavra era mencionada em seu objeto e referenciavam "Servidor público" ou buscas por menções as práticas de sustentabilidade ESG resultaram em licitações relacionadas a saneamento básico de "**Esg**oto". Estes casos foram tratados por supervisão e removidos manualmente e, em alguns casos que geraram somente *outliers* ou com frequência muito maior que positivos*,* a *keyword* foi removida dos vetores de busca.

3.4 Descrição do experimento

Inicialmente, foi feito um levantamento das plataformas e portais de B2B disponíveis no mercado brasileiro e base de conhecimento tácito do time que já atua há mais de 10 anos apoiando as transações eletrônicas de clientes somando-se também uma busca por interesse de novas empresas em diversos segmentos, incluindo os que já eram conhecidos e outros segmentos de mercado que se tinha interesse em ingressar. Em seguida, as plataformas foram elencadas em 3 grupos gerais:

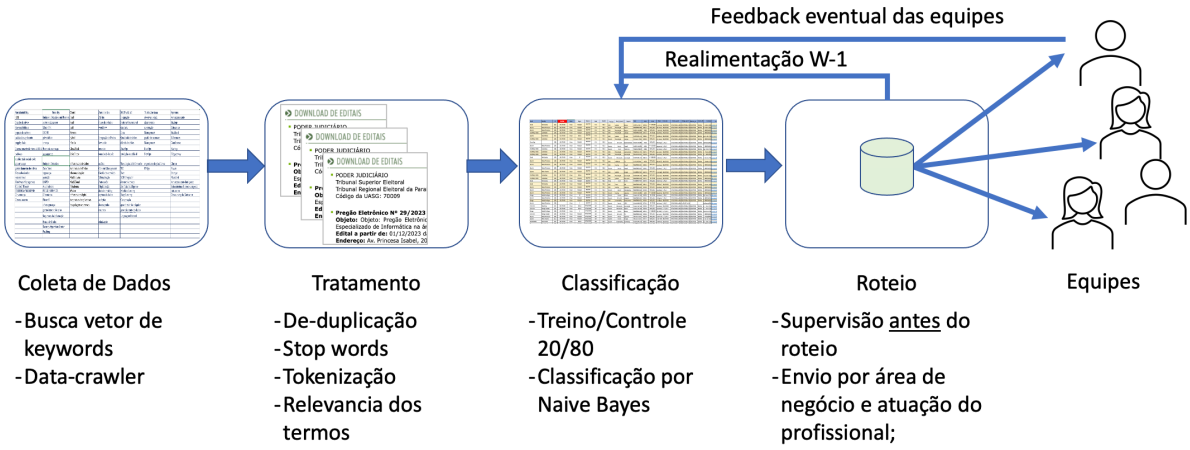
1) Plataforma própria: desenvolvida *in-house* ou adquirida pela empresa.
2) Plataformas de mercado com foco em cliente único: Petronect (Petrobras), GEP (Telefónica), Portal Licitar Digital (Cemig), entre outros.
3) Marketplace/ePV: empresa com o objetivo de conectar cliente comprador a empresas fornecedoras. Alguns exemplos: SAP Ariba, Mercado Eletrônico, Coupa, Nimbi, Aquanima, Compras.gov.br entre outros.

Com o objetivo de desenhar um piloto funcional dentro do prazo estabelecido de 6 meses decidiu-se pelo grupo (3) por permitir que o esforço de pesquisa de uma implementação do modelo pudesse gerar um ganho de escala e tivesse atuação em múltiplos clientes que operassem em uma plataforma. Esperava-se que, uma vez estabelecido um processo, este pudesse ser expandido observando-se alguns ajustes a serem detalhados a posteriori. Desta forma optou-se pelo portal Compras.gov.br (passo a referenciar como Portal), pois este permite acesso a todos os processos de compras das entidades governamentais no Brasil, incluindo-se os das esferas municipais, estaduais, federais e órgãos relacionados, tais como: polícias, forças armadas, institutos de pesquisas e preservação, entre outros.

A partir do levantamento inicial no Portal selecionado, foi identificado que ele permite buscas em qualquer ponto no tempo, mas limita a uma janela máxima de busca de 15 dias corridos (2 semanas). Desta forma, definiu-se uma frequência inicial de busca semanal. Para validar se as oportunidades têm aderência ao portfólio da empresa, foi realizada uma busca qualitativa inicialmente de forma manual no portal e os dados obtidos foram armazenados em planilha, onde buscava-se capturar os dados básicos para a classificação da oportunidade, ainda sem o link de acesso direto.

Este processo foi repetido por algumas semanas. Para assegurar a assertividade do processo, as oportunidades identificadas foram compartilhadas com a área de vendas e os profissionais que receberam o material foram consultados, confirmaram o valor de receber regularmente e concordaram em contribuir para a supervisão das informações fornecendo suas percepções que seriam consideradas na realimentação da base de treinamento da semana seguinte.

Figura 3 Ciclo semanal de tratamento de dados



Fonte: O autor

Após validado que esta extração tinha valor para as equipes comerciais, foi feita a implementação da automação através do uso do *crawler*. O Playwright (https://playwright.dev/) como crawler foi parametrizado para navegação até a página de busca e depois a repetição de ciclos de busca por cada lista de termos por área de negócio ou coluna. Inicialmente para se acompanhar a visualização da navegação das páginas para coleta como uma supervisão visual e à medida que o processo foi sendo internalizado após algumas semanas, passou-se a usá-lo na versão sem visualização (ou *headless*). Os dados capturados nesta etapa foram salvos em um arquivo formato XLSX para futuro processamento.

Coletados os dados básicos de uma licitação, pensou-se em tentar obter o documento de publicação da licitação na íntegra que fica disponível através de um link em cada processo e incluí-lo no processamento. Contudo este fica protegido de acesso automatizado usando o sistema de *captcha*, muito comumente abordado em sites onde o usuário passa por um desafio de interpretar em imagens, objetos ou caracteres alfanuméricos, acertando a interpretação o conteúdo da íntegra é disponibilizado para download.

Após análise, inclusive com uma indicação de um serviço pago na internet que oferece a automação de identificação de *captcha*, optou-se por concentrar os esforços utilizando-se o campo "Objeto" da licitação como o texto para análise de máquina, pois mesmo que se obtivesse de uma forma razoável, isso geraria um volume e complexidade de análise de linguagem natural que fugiria do escopo do projeto, desviando de uma possível abordagem de implementação prática.

Nesta etapa executa-se a fase de tratamento, que inclui pré-processamento (deduplicação, falsos positivos, remoção de stop words e tokenização) como já descrito.

Na etapa seguinte de classificação, usa-se a base histórica de licitações supervisionada até a semana anterior (W-1), para treinar e testar o modelo multinomial de Naive Bayes adotado, em uma proporção de 20% treino e 80% validação. Na prática a cada semana essa base é incrementada em torno de 20 oportunidades que foram supervisionadas na semana anterior, incluindo-se os comentários das equipes sobre os dados que receberam.

A forma de supervisionamento ocorre com a classificação da Área de negócio oriunda do vetor (da lista de *keywords*) >> Área prevista pelo modelo >> Área final supervisionada. Ao retroalimentar o modelo na semana seguinte ele vai considerar a relação entre o objeto da licitação e a classificação da Área supervisionada. As classificações intermediárias foram mantidas como base de dados para eventual análise futura para ajuste fino do modelo adotado e/ou revisão de implementação de outras implementações possíveis de modelos de NLP.

Após a classificação, as oportunidades são direcionadas para as equipes comerciais que analisam, validam se estão dentro do escopo de portfólio e conectam as equipes necessárias para desenvolver as propostas ao seguindo seus fluxos de trabalho.

O *feedback* dos dados enviados, quanto à aderência ao portfólio quanto elegibilidade é feita de 2 formas: 1) diretamente na planilha compartilhada em uma coluna de comentários e 2) Contato ativo com as equipes buscando questionar a validade em forma de amostra algumas oportunidades que se julgue de maior probabilidade de acerto (no processo de classificação).

3.5 Resultados e análises

A automação do processo de captura dos dados resultou em um volume muito maior de informações tabuladas, em comparação com o processo manual. Como o volume aumentou semanalmente, ficou evidente a crescente necessidade de implementação de um processo de limpeza e de-duplicação dos dados. Neste aspecto houve um ganho no volume geral de licitações coletadas semanalmente que saltou de dezenas para centenas semanalmente.

A classificação da área de negócio usando-se Machine Learning resultou em uma taxa de acertos em torno de 58% que variou de acordo com ajustes de treinamento e que segue sendo ponto de estudo durante a execução semanal.

O modelo multinomial de NaiveBayes se mostrou adequado inicialmente, porém algumas análises estão sendo feitas no algoritmo como um todo, desde o pré-processamento até o tratamento, pois acredita-se que pode estar influenciando a classificação e acurácia final.

Por fim, apurando-se de forma prática, até a presente data, o processo foi executado experimentalmente durante a implementação durante o terceiro trimestre de 2023 e desde o

início do quarto trimestre de 2023 formalmente, enviando-se os dados coletados às equipes de trabalho, durante 10 semanas consecutivas.

Em torno de 3500 oportunidades foram capturadas pela busca do *crawler*, dessas 900 resultaram do processo de tratamento reduzindo-se a 271 que foram classificadas, gerando um aumento de 75% no volume médico semanal de oportunidades tratadas (*versus* processo inteiramente manual) e ficaram na base acumulada de treino/supervisão, destas 3 foram identificadas como aderentes, sendo que 1 delas foi de fato trabalhada por uma empresa parceira com valor estimado de USD 20k. Este resultado inicial, produziu retornos positivos das equipes, o que encoraja a continuar a busca pela otimização do projeto e agregação de novos portais, visto que representa um potencial novo de negócios que não estavam sendo capturados.

Devido a base inicial de treinamento ainda em construção, um ponto que se imagina como evolução da classificação à medida que a base supervisionada crescer, podendo ser testado outros modelos e comparado quanto a taxa de acerto e eventualmente substituí-lo.

Apesar do modelo poder atuar de forma não supervisionada, mostra-se importante, tanto do ponto de vista de ajuste fino do modelo de ML quanto de proximidade com as equipes comerciais mantendo-se o projeto o mais próximo da realidade de negócio, evitando a geração de resultados puramente teóricos e sem aplicação rápida no ambiente de negócio.

Uma questão importante que se identificou durante a implementação é a correta comunicação dos resultados obtidos pela classificação de ML às equipes comerciais, para que avaliem as oportunidades dentro de um prazo exequível e para que forneçam comentários e insumos para se ajustar o modelo, tais como: palavras-chaves mais bem sucedidas, variações de palavras que levam a falsos positivos etc. Com a futura evolução e ganho de escala do modelo, poderá ser interessante uma conexão com a plataforma de gestão de oportunidades da empresa e a base do modelo, a fim de automatizar a retroalimentação do modelo quanto a oportunidades bem-sucedidas que venham a converter em negócios fechados, evitando-se assim a dependência da interação humana nesta parte de alimentação do treinamento.

Como parte do processo de estudo e exploração do portal Compras.gov.br, identifica-se já a necessidade de manter o projeto atualizado brevemente devido à implementação de nova lei de licitações 14.133/21 que vai demandar análise na interface de busca e eventual ajuste no *data-crawler*, espera-se que ajustada essa etapa o resto do processo, como tabulação, classificação e roteiro, se mantenha inalterado, desde que seja possível capturar o mesmo nível de informações das licitações originalmente, o que se espera que aconteça dado que a natureza descrição das licitações não se altera.

Analogamente, pensando-se em expandir o modelo para outros portais, imagina-se que a seguinte abordagem deve ser seguida:

1) Exploração do potencial de negócio do novo portal a ser agregado;
2) Avaliação da complexidade no acesso ao portal (processo de autenticação) quanto a automação;
3) Exploração da interface e similaridade de nível de informação disponível;
4) Construção modular o data-crawler correspondente ao portal específico;
5) Disponibilização dos dados para o passo de tratamento e classificação já estabelecidos.

Com a agregação de vários portais heterogêneos ao modelo, alguns pontos de pesquisa e discussão poderão ser necessários, tais como:

1) As licitações aglomeradas poderão ser consolidadas em um *datalake*, tanto para treinamento quanto histórico e alguma tecnologia de banco de dados implementada ao invés de simples tabulação dos dados, facilitando a manutenção dos dados;
2) Como portais de empresas privadas não têm obrigatoriedade da lei da transparência, estas oportunidades podem não estar disponíveis publicamente e ser acessíveis somente a perfis cadastrados no portal, para amplo acesso das equipes aos documentos das licitações, eventualmente será necessário que no momento da captura do edital/licitação a íntegra do documento seja armazenada localmente ou em sistema próprio para que o link ao conteúdo possa ser distribuído entre as equipes de trabalho;
3) Existem oportunidades de melhoria da busca e classificação em si quanto da portabilidade do mesmo método para outros portais.

Um ponto de evolução futura do projeto na parte de tratamento e classificação por ML que se capture a íntegra do processo e não somente seu resumo/objeto fazendo-se um treinamento sobre todo o conteúdo do texto do edital. Aqui abre-se uma possibilidade muito grande de formas de se tratar o texto, inclusive novas estratégias usando-se outros modelos de AI dado que o volume de dados para treinamento daria um salto quantitativo.

Uma implementação que pode se tornar bastante interessante do ponto de vista de valorização do projeto é se trabalhar uma forma visual de monitorar os dados e processo, sob formato de infográfico ou *dashboard* que possui apelo de negócio grande entre os líderes executivos da organização, visto que pode ser acompanhado ao longo dos trimestres e oferecer informações de comparação de crescimento e funil de vendas período sobre período (ex. ano contra ano, trimestre contra trimestre ou subsequente *versus* anterior).

4 CONCLUSÃO

A implementação do modelo mostrou-se viável de baixa complexidade de implementação, o que torna interessante do ponto de vista de negócio. Com uma base pequena de treinamento e validação, o classificador de Naive Bayes se mostrou razoavelmente adequado, mas deve seguir-se a avaliação de seus parâmetros a medida que a base supervisionada de treino aumenta e futuramente outros modelos deveriam ser avaliados em busca de maior acurácia e menor necessidade de supervisão, inclusive para outras etapas do processo como pré-processamento e tratamento.

REFERÊNCIAS

Simões, H. & Jorquera, G. (2021) O impacto da Tecnologia no consume pós-pandemia. https://gente.globo.com/o-impacto-da-tecnologia-no-consumo-pos-pandemia/ O Globo.

Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of Web crawlers for information retrieval. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1218.

Mercado Eletrônico. (2023) fonte: https://blog.me.com.br/pt/e-procurement-gestao-compras

Rauen, A. T. O. (2022). Compras públicas para inovação no Brasil: Caracterização dos contratos de compras públicas existentes no sistema integrado de administração de serviços gerais.

Shanks, J. (2016). Social Selling mastery. John Willey & Sons, Inc

Singh, M. S. A. D. J., & Varnica, B. (2014). Web crawler: Extracting the web data. *International Journal of Computer Trends and Technology*, *13*(3), 132-137.

Thelwall, M. (2001). A web crawler design for data mining. *Journal of information science*, *27*(5), 319-325.

Webb, G. (2016). Naïve Bayes. Monash University, Australia
https://www.researchgate.net/publication/306313918

Yang, F. J. (2018, December). An implementation of naive bayes classifier. In *2018 International conference on computational science and computational intelligence (CSCI)*(pp. 301-306). IEEE.

Zhang, H. (2004). The optimality of naive Bayes. *Aa*, *1*(2), 3.

Apêndice A – Lista de keywords para buscas

Sustaintability	Security	Cloud	Automation	DATA & AI	Turbo/Instana	Storage
Gestão de ativos	Endpoint Detection and Response (EDR)	PaaS	AIOps	observabilidade	Armazenamento	Renew
Sustentabilidade	controle de acesso	SaaS	fluxo de trabalho	desempenho	Backup	Revenue
pegada de carbono	MaaS 360	IaaS	workflow	data lake	automação	DataCenter
cadeia de suprimento	GDPR	Nuvem	RPA	Etica	application Resource Management (ARM)	Backlevel
supply chain	privacidade	hybrid	Integração de Dados	Qualidade de dados	gestão de recursos	Kubernetes
ESG	ameaça	hibrida	low code	ciencia de dados	Application Performance Management (APM)	Contêineres
Emissions Management	Resposta a ameaça	CloudPack	no code	DataOps	DevOps	Sterling
Ambiental	ransomware	SalesForce	tomada de decisão	Inteligencia Artificial	engenharia de plataforma	FileGateway
Descarbonization	detecção	Infraestrutura	ECM	Tecnologia da Informação	ITOps	Cluster
Assessment	Zero Trust	Arquitetura	Content Management	TIC	NettOps	Power
ativos	segurança	telecomunicações	Gestão de conteúdo	Data		SYSTEMS
Asset	proteção	Multinuvem	Virtualização	B2B Integrator		Massload
Governança	LGPD	MultiCloud	Automação	desenvolvimento		
Gerenciamento	Antimalware	Plataforma	Digitalização	Artificial Intelligence		
Hardware Management Console (HMC)	FIELD SERVICE	Walker	decision making	Machine Leaning		
Environmental	Cibernética	BUSINESS SUPPORT	data capture	Deep Leaning		
Control Tower	Detection	ITOC	analytics	Computação		
TRIRIGA MAXIMO	Management Detection and Response (MDR)		análise	Integração		
	Guardium			SAP		
	DataCenter			LGPD		
	Banco de Dados					
	SCC					
	Patching					