UNIVERSIDADE PRESBITERIANA MACKENZIE

AMANDA SCHWANTES MAIA

ELAINE THAMIRES PEREIRA

GIOVANA VIEIRA DA SILVA

IAGO FONSECA MARINHO PEREIRA

MURILO GUIMARAES

ANÁLISE DE AVALIAÇÕES SOBRE VINHOS ATRAVÉS DO PROCESSAMENTO DE LINGUAGEM NATURAL

AMANDA SCHWANTES MAIA TIA: 10922519963 ELAINE THAMIRES PEREIRA TIA: 10922517650 GIOVANA VIEIRA DA SILVA TIA: 10922520791 IAGO FONSECA MARINHO PEREIRA TIA: 10922500138

MURILO GUIMARAES TIA: 10922505458

ANÁLISE DE AVALIAÇÕES SOBRE VINHOS ATRAVÉS PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de disciplina intitulada Projeto Aplicado II, componente do curso Ciência de Dados da Universidade Presbiteriana Mackenzie.

Orientador: Prof. Anderson Adaime De Borba

AMANDA SCHWANTES MAIA ELAINE THAMIRES PEREIRA GIOVANA VIEIRA DA SILVA IAGO FONSECA MARINHO PEREIRA MURILO GUIMARAES

ANÁLISE DE AVALIAÇÕES SOBRE VINHO	OS ATRAVÉS DO PROCESSAMENTO
DE LINGUAGEM NATURAL	

Trab	alho	de	Conclus	ão c	le	discipl	ina
intitu	ılado F	roje	to Aplica	do II,	cor	npone	nte
do	curso	• (Ciência	de	Da	dos	da
Universidade Presbiteriana Mackenzie.							

BANCA EXAMINADORA	

Aprovado em:

RESUMO

Este projeto acadêmico tem como intuito desenvolver um ponto de vista da ciência de dados através de conhecimentos obtidos por meio de disciplinas como Análise Estatística Preditiva, Aprendizado de Máquina, Aquisição e Preparação de Dados. A ideia é utilizar e relacionar tais aprendizagens na prática ao elaborar esse estudo. Além disso, deve ser projetado e implementado ao projeto a manipulação de imagens ou textos.

Nesse contexto, o trabalho envolve a análise de sentimentos das pessoas com base em avaliações da *web* relacionadas ao mundo dos vinhos. À vista disso, foi criada uma empresa fictícia de tecnologia, a qual oferecerá ao cliente J2Winner uma pesquisa de mercado, isto é, uma lista de vinhos para que o mesmo possa investir de maneira precisa. A empresa que receberá o serviço atua como concorrente de um grande site de vendas de vinhos. O objetivo é compreender as avaliações dos consumidores sobre os diferentes rótulos disponíveis *on-line*, para analisar, posteriormente, se os mais bem avaliados são também os mais vendidos e, com isso, o empreendimento identificaria quais garrafas teriam maior potencial comercial.

Para a extração dos insights ditos acima, será empregue técnicas de análise exploratória de dados, estudo da linguagem natural (*NLP*) e aprendizado de máquina e, além disso, algumas etapas essenciais estão incluídas como a coleta de dados, o pré-processamento textual, a modelagem dos dados, análise de correlações, dentre outras. Contudo, o projeto busca reconhecer padrões das opiniões dos clientes sobre vinhos e, dessa forma, permitirá que a J2Winner tome decisões sobre quais rótulos devem ser incluídos em seu catálogo, visando fornecer aos clientes os vinhos mais bem avaliados e satisfatórios, e consequentemente, melhorando sua estratégia de compras e expandir a satisfação do cliente.

Palavras-chaves: Análise de Sentimentos, Avaliações da *Web*, Vinhos, Aprendizado de Máquina, Exploração de Dados, Tomada de Decisões.

ABSTRACT

This academic project aims to develop a perspective on data science through knowledge acquired from disciplines such as Predictive Statistical Analysis, Machine Learning, Data Acquisition, and Preparation. The idea is to use and relate such learnings in practice while conducting this study. Additionally, image or text manipulation should be designed and implemented as part of the project.

In this context, the work involves analyzing people's sentiments based on web reviews related to the world of wines. To this end, a fictional technology company has been created, which will offer J2Winner clients a market research service, namely, a list of wines for them to invest in precisely. The company receiving the service operates as a competitor to a major online wine sales website. The objective is to understand consumer reviews of different available labels online, and subsequently analyze if the highest-rated ones are also the best-sellers. Through this, the business would identify which bottles have the highest commercial potential.

To extract the insights mentioned above, techniques such as exploratory data analysis, natural language processing (NLP), and machine learning will be employed. In addition, essential steps are included, such as data collection, textual preprocessing, data modeling, correlation analysis, among others. However, the project aims to recognize patterns in customer opinions about wines and, thus, enable J2Winner to make decisions about which labels should be included in its catalog. This aims to provide customers with the highest-rated and most satisfactory wines, consequently improving their purchasing strategy and enhancing customer satisfaction.

Keywords: Sentiment Analysis, Web Reviews, Wines, Machine Learning, Data Exploration, Decision-Making.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1.	INTRODUÇÃO		
2.	MET	ODOLOGIA	11
	2.1.	COLETA DE DADOS	11
	2.2.	PRÉ-PROCESSAMENTO DOS DADOS	12
	2.3.	ANÁLISE DE SENTIMENTO	13
3.	OBJ	ETIVOS DO PROJETO	14
	3.1.	METAS	15
	3.2.	RESULTADOS ESPERADOS	15
4.	DEF	INIÇÃO DA ORGANIZAÇÃO	16
5.	DA E	EXTRAÇÃO DE DADOS	16
6.	CRO	NOGRAMA DE ATIVIDADES	17
7.		SIDERAÇÕES FINAIS	
8.	REF	ERÊNCIAS BIBLIOGRÁFICAS	20
9.	GLO	SSÁRIO	21

1. INTRODUÇÃO

A história do vinho e seu advento circundam em muitas lendas e, por isso, estudiosos dizem que não se sabe ao certo como a bebida foi criada. Porém, segundo o arqueólogo *Patrick Edward McGovern*, que escreveu o livro "Ancient *Wine*", há indícios moleculares da presença de vinho entre 3.000 e 6.000 a.C. e, outros, ainda dizem que sua provável origem é do Oriente Médio. Já alguns, declaram que tenha sido desenvolvido ao acaso, deixaram as uvas amassadas num frasco e elas fermentaram espontaneamente.

Sob o contexto alimentar, o que se sabe é que no passado o vinho estava presente em diversos momentos da refeição e era apenas um alimento à mesa como outro qualquer, porém, ao longo dos anos o manejo mudou, as técnicas tornaram-se mais engenhosas e, consequentemente, passou a ser um elemento ímpar num contexto social. Com isso, gradualmente, o próprio consumidor reagiu a este aperfeiçoamento, buscando novos aromas, sabores e novas circunstâncias para inseri-lo (De Almeida, 2020).

Na ótica religiosa, Noé foi seu inventor. Ao atracar sua arca na região de Monte Ararat, hoje intitulada como Turquia, plantou vinhas e com as uvas, ele fez vinho, entretanto, de modo infeliz, foi flagrado por seus filhos, despido e caído no chão (Gênesis 9:20-25). Posteriormente a isso, em Caná da Galileia, durante um casamento, o vinho acabou, algo que era considerado uma adversidade para época, então Jesus, após atender um pedido de Maria, sua mãe, transformou a água em vinho, eis então, ao compadecer-se àquele casal, o primeiro milagre de Cristo (João 2:1-11).

Nessa lógica, ainda num entendimento espiritual, na Bíblia, além desses capítulos, o vinho desempenha uma conotação simbólica, tendo significados que variam conforme o enredo. Em certas passagens, ele representa alegria e abundância, celebração e comunhão e até simboliza a singularidade religiosa onde, em algumas interpretações, traduz-se como sangue de Cristo. Todavia, também é associado ao excesso e à embriaguez, como vivenciado por Noé, e à perdição, salientando os perigos e vícios do homem. Assim, sendo estimado como salvação ou perdição, sabe-se que o vinho possui papel essencial na narrativa da vida

mundana ao longo do tempo, independentemente da perspectiva que esteja inserido.

Além dessas e outras narrativas, nota-se a vasta história que o cerca, tendo uma carga cultural enorme. Como mencionado no texto, sua ascensão reflete transformações no corpo social e em seus costumes, tornando-o num objeto de estudo atraente. Sendo assim, diante de tanta pluralidade, ao analisar o sentimento das pessoas em relação a essa bebida será possível descobrir novas percepções sobre ela, uma vez que desencadeou um papel íntegro na cultura humana, desde sua origem complexa até sua evolução popular e cerimonial que se tem hoje.

Além do que, ao analisar sentimentos, contextualizando com o objetivo do projeto, será possível identificar as preferências dos consumidores e suas opiniões enquanto aos sabores e marcas, as tendências de mercado e o modo como o vinho é apreciado em diferentes cenários. Com isso, essas informações poderão ser úteis para as redes vinícolas, produtores e vendedores, como o caso do estudo vigente, permitindo-lhes estipular estratégias segundo as demandas e expectativas dos clientes, tornando a comercialização mais estruturada e, portanto, podendo ter um aumento comercial significativo.

Entender a experiência do consumidor é crucial para a exploração dos sentimentos num estudo como esse. Compreender como ele se sente em relação ao vinho, quais notas, qual sua tonalidade, corpo, tanino, acidez, complexidade, persistência, entre outros diversos aspectos que podem surgir ao avaliar um vinho, são indispensáveis no momento da elaboração do modelo utilizando o Processamento de Linguagem Natural (*NLP*). Ainda, além dessas características específicas, os adjetivos comuns, como "bom" e "ruim", "ótimo" e "péssimo", ou até mesmo expressões como "não gostei" e "não recomendo", também serão usadas para o treinamento do algoritmo.

Ao coletar avaliações de vinhos disponíveis na *internet* que contenham esses adjetivos e expressões, eles se tornam dados relevantes para treinar o modelo, uma vez que refletem a perspectiva real do público e contribuem para uma avaliação rigorosa e objetiva. Com o auxílio dessas opiniões, o algoritmo pode aprender a distinguir entre avaliações positivas, neutras e negativas, identificando detalhes sutis nas palavras e expressões publicadas pelos clientes.

Em síntese, ao aproximar o projeto em si com o que foi dito acima e os *insights* que ainda serão coletados através das técnicas de exploração de dados, pode-se estabelecer correlações interessantes entre a história, a cultura e as perspectivas modernas, dessa maneira, refletindo o encargo do vinho na sociedade, bem como, evidenciando seu crescimento e adaptação à medida do tempo. Portanto, realizar um projeto como esse não promove apenas oportunidades comerciais, mas também proporciona a contemplação, concepção e conservação dessa parte essencial da civilização mundial.

2. METODOLOGIA

No contexto atual, as avaliações dos produtos e serviços representam muita relevância para destacar determinados produtos dos demais. Além do item ser exibido prioritariamente nas buscas, novos clientes costumam buscar referências antes de efetuar uma compra. Os comentários dos avaliadores dão evidência a uma opinião, quais características daquele artigo agradaram os consumidores. Do mesmo modo, muitas vezes expressam um sentimento de satisfação ou insatisfação.

A análise de sentimentos desempenha um papel crucial na compreensão das opiniões dos consumidores em relação aos produtos. Nesta etapa, descreveremos a metodologia que será aplicada para realizar uma análise de sentimento das avaliações de vinhos no site Mercado Livre. Este processo nos permitirá extrair informações valiosas sobre a satisfação do cliente, melhorar a qualidade dos produtos oferecidos, bem como desenvolver estratégias de marketing mais eficientes.

2.1. COLETA DE DADOS

O objetivo da etapa de coleta de dados é obter as avaliações dos clientes que foram realizadas no site Mercado Livre. Assim, teremos acesso a um conjunto de dados que irá fundamentar nossa análise nas etapas seguintes. O processo de coleta de dados será realizado através da técnica de "Web Scraping", o qual se refere ao ato de extrair informações diretamente das páginas web.

Para executar esse processo com eficiência, utilizaremos bibliotecas do *Python* específicas, como, por exemplo, "*BeautifulSoup*" e "*Scrapy*". Tais bibliotecas serão configuradas de modo a automatizar o processo, algo que seria muito custoso caso fosse realizado manualmente.

Primeiramente, identificamos as seções de páginas *web* dedicadas aos vinhos. Através disso, buscaremos os itens que possuem informações potencialmente relevantes e com maior riqueza de dados. Após mapear essas

páginas, utilizaremos a biblioteca "*Requests*" para enviar solicitações *HTTP*, permitindo acessar os conteúdos dessas páginas.

Em posse dos conteúdos, utilizaremos a biblioteca "BeautifulSoup" para analisar a estrutura da página. Isso nos permitirá identificar os elementos HTML que contêm as avaliações, como elementos de parágrafos ou classificações por estrelas. Em seguida, pretende-se extrair o texto das avaliações, bem como qualquer outra informação útil, como a data da avaliação, a classificação do vinho e o nome do consumidor.

Os dados serão coletados e armazenados em um arquivo *CSV*, para poder ser acessado facilmente nas etapas subsequentes. É importante evidenciar que essa coleta de dados será realizada de acordo com as diretrizes de uso do site, evitando sobrecarregar os servidores. Além disso, todo e qualquer dado obtido nesse processo, será anonimizado para respeitar a Lei Geral de Proteção de Dados.

2.2. PRÉ-PROCESSAMENTO DE DADOS

Posteriormente à coleta de dados, é fundamental manipular os textos das avaliações com o propósito de permitir e otimizar o processo de análise. Para isso, é importante remover caracteres especiais, números e pontuações que não são relevantes para análise de sentimento.

Para auxiliar na etapa de pré-processamento de dados utilizaremos a bibliotecas especializadas, como a "NTLK", a qual oferece recursos para processamento de linguagem natural, incluindo tokenização e remoção de "stop words".

A tokenização é o processo de dividir determinado texto em unidades menores chamadas "tokens". Em linguística e no processamento de linguagem natural, um "token" geralmente representa uma unidade semântica que faz sentido em um contexto linguístico. Este é um método fundamental no pré-processamento de texto antes do treinamento de modelos de linguagem.

A remoção de "stop words" também é imprescindível, pois são palavras que ocorrem com muita frequência no texto, mas geralmente são irrelevantes, pois não contribuem significativamente para a análise de sentimento. Alguns exemplos são: e, em, do, um, por, com, etc.

Por último, é importante realizar a normalização do texto, assim, padronizando letras maiúsculas e minúsculas. Esse processo é importante, pois elimina duplicações, por exemplo, as palavras "BOM", "Bom" e "bom" serão tratadas como uma única palavra e não como palavras distintas. A normalização garante a consistência da análise, evitando a contagem ou tratamento separado de palavras que são idênticas.

2.3. ANÁLISE DE SENTIMENTO

A fase da análise de sentimento é o âmago do projeto, nela serão utilizados artifícios e métricas de aprendizado de máquina para determinar os sentimentos relacionados às avaliações dos vinhos. Converter os textos não estruturados em vetores numéricos será essencial, pois os algoritmos de aprendizado de máquina geralmente trabalham com números. Executando a vetorização do texto, os dados serão traduzidos e aptos para serem compreendidos por bibliotecas especializadas, como o caso da "Scikit-learn". Esta biblioteca oferece algoritmos que podem ser utilizados para análise de sentimentos, "Naive Bayes" e "Support Vector Machines" são exemplos.

A conversão de textos para vetores numéricos também ajuda na eficiência computacional, já que reduzem a dimensionalidade, facilitam os cálculos matemáticos, como também facilita a captura de relações semânticas, já que palavras semelhantes tendem a ter representações vetoriais próximas no espaço vetorial.

Com a organização desta estrutura, avançaremos para o treinamento do modelo. Um conjunto de dados são rotulados, podem ser atribuídos adjetivos como "positivo", "negativo" ou "neutro". O modelo aprende através desses rótulos, se adequando para identificar padrões que os representem. O objetivo principal desta

etapa é ensinar um modelo de aprendizado de máquina a reconhecer automaticamente esses sentimentos nas avaliações não rotuladas.

Durante esse treinamento, é importantíssimo avaliar o desempenho do modelo, utilizando métricas que validem a precisão do modelo. Um exemplo dessas métricas é o *F1-Score*, ele é particularmente útil em cenários onde a distribuição das classes não é uniforme e onde é necessário encontrar um equilíbrio entre a identificação correta de exemplos positivos e a minimização de falsos positivos e falsos negativos.

Para concluir a análise de sentimento, o modelo treinado será aplicado às avaliações dos vinhos. Conforme os padrões assimilados durante o treinamento, o modelo irá categorizar as opiniões dos consumidores em cada rótulo que foi definido ("positivo", "negativo" e "neutro"). Isso possibilitará uma análise automatizada, permitindo *insights* e que conhecimentos valiosos sejam extraídos para uma melhor compreensão da percepção do público em relação aos vinhos.

3. OBJETIVOS DO PROJETO

O principal objetivo é realizar uma análise abrangente do comportamento dos consumidores de vinho por meio de um processo estruturado. Para alcançar isso, estabelecemos metas claras e passos específicos a serem seguidos.

Em primeiro lugar, planejamos coletar dados valiosos dos consumidores, concentrando-nos nos comentários disponíveis no site do Mercado Livre. Utilizaremos técnicas de "Web Scraping" para extrair esses dados de maneira eficiente, garantindo um conjunto de informações robusto e representativo.

Em seguida, entraremos na fase de compreensão do comportamento do consumidor. Aqui, faremos uso do aprendizado de máquina para analisar os dados coletados. Nossa intenção é desvendar tendências, preferências e padrões que os consumidores demonstram em relação aos diferentes tipos de vinhos, marcas e preços. Essa análise nos permitirá obter *insights* valiosos sobre o mercado de vinhos e o que motiva suas escolhas.

Uma parte crucial do nosso projeto é a classificação dos comentários dos consumidores. Utilizaremos nosso modelo de aprendizado de máquina treinado com

os dados coletados para determinar se os comentários são positivos ou negativos. Isso nos permitirá avaliar a satisfação dos consumidores com os vinhos e identificar áreas de melhoria para a J2Winner.

3.1. METAS

Nossa meta em relação à classificação é alcançar uma alta precisão, garantindo que nossas avaliações sejam confiáveis e significativas. Estabeleceremos um valor específico para essa métrica, garantindo que nosso modelo seja altamente confiável na diferenciação entre comentários positivos e negativos.

Com base nas classificações obtidas, será gerado uma lista de recomendações de vinhos. Nela, destacaremos o vinho melhor avaliado no momento, fornecendo à J2Winner uma sugestão sólida e confiável.

Em resumo, nosso projeto visa coletar dados, compreender o comportamento do consumidor, classificar comentários e gerar recomendações de vinhos com base em análises específicas.

3.2. RESULTADOS ESPERADOS

A expectativa é que, por meio do uso de técnicas avançadas de Machine Learning, possamos realizar uma análise aprofundada dos comentários dos clientes. Essa análise será uma ferramenta valiosa para avaliar a qualidade dos vinhos que as empresas que fazem parte do Mercado Livre oferecem, bem como para compreender as opiniões sobre aspectos fundamentais, como sabor, aroma, corpo e outros aspectos sensoriais. Tais informações são vitais para aprimorar continuamente a seleção de vinhos da empresa J2Winner, garantindo que seus produtos atendam plenamente às expectativas dos consumidores.

Além disso, essa análise nos proporcionará uma visão competitiva do mercado, nos ajudando a identificar os pontos fortes e fracos de cada marca em relação à concorrência, fornecendo um suporte consistente para tomadas de decisão estratégicas.

Com base nos insights derivados das preferências dos clientes, poderemos

desenvolver estratégias de marketing personalizadas. Essa abordagem implica recomendar vinhos específicos com base no histórico de avaliações dos consumidores, aumentando assim a relevância das ofertas para cada cliente e fortalecendo o relacionamento deles com a J2Winner.

Adicionalmente, ao identificar problemas recorrentes mencionados nos comentários, como questões relacionadas à entrega ou ao atendimento ao cliente, estaremos em posição de tomar medidas direcionadas com base nessas informações. Isso não apenas melhora a experiência do cliente, mas também solidifica a fidelidade à marca, estabelecendo uma relação de confiança e satisfação duradoura com os consumidores.

4. DEFINIÇÃO DA ORGANIZAÇÃO

Nosso estudo de caso será a empresa J2Winner, a qual é especializada na venda de vinhos de diversas marcas e nacionalidades, operando em um ambiente exclusivamente *on-line* (j2winner.com.br). Sua missão é oferecer aos clientes uma seleção diversificada de vinhos de alta qualidade de diferentes regiões do mundo, proporcionando uma experiência de compra única e personalizada e, apesar de ser relativamente nova no ramo, a organização tem como objetivo estratégico a curto prazo o crescimento no mercado de vinhos *on-line* e o aprimoramento contínuo da personalização das recomendações de produtos.

5. DA EXTRAÇÃO E FONTE DOS DADOS

O principal objetivo deste projeto acadêmico é desenvolver um estudo utilizando técnicas de *Machine Learning* para a mineração de opiniões e análise de sentimentos em avaliações de vinhos disponíveis no site mercadolivre.com.br. Trata-se de um *e-commerce* amplamente conhecido e procurado devido sua variação de produtos e marcas e, ainda, que conta com sistema de avaliações em cada produto. Com base nas características de produtos que recebem opiniões por escrito de consumidores, é possível gerar com análise destes textos uma classificação positiva ou negativa, o que motiva as empresas a investirem em determinados produtos para um determinado público, neste caso, consumidores de vinhos. Com base nesses *insights*, o projeto visa criar um ranking dos vinhos mais

bem avaliados e comentados e gerar recomendações de produtos para a equipe de vendas da J2Winner, aprimorando assim a experiência de compra do consumidor e impulsionando as vendas.

Os dados que servirão como base para este estudo consistem em avaliações em texto de vinhos publicadas nesta plataforma de vendas. Essas avaliações incluirão informações relevantes, como a descrição dos produtos, as classificações dos consumidores e o nosso foco que são os comentários em textos deixados pelos compradores com opiniões. Para coletar esses dados, utilizaremos uma técnica chamada "Web Scraping".

O "Web Scraping" é um método de coleta de informações na web que envolve a extração de dados de páginas da web, incluindo aquelas com conteúdo dinâmico gerado por usuários, como comentários. Com a ajuda da linguagem Python e bibliotecas específicas, será possível acessar e extrair as avaliações e comentários dos vinhos de forma automatizada e, a partir disso, utilizar os métodos aplicados para Machine Learning.

A coleta e análise desses dados por meio do "Web Scraping" serão essenciais para cumprir os objetivos propostos neste projeto, permitindo-nos extrair informações relevantes e realizar análises de sentimentos em um conjunto de dados abrangente e representativo das avaliações de vinhos disponíveis na plataforma.

6. CRONOGRAMA DE ATIVIDADES

Até 11/09/2023:

Definição do grupo de trabalho;

Definição das premissas, proposta (texto ou imagem);

Distribuição de atividades;

Objetivos, metas e milestones;

Cronograma de atividades;

Definição da empresa e área de atuação;

Pesquisa de benchmark e referências;

Consolidação de entrega.

Até 02/10/2023:

Aprofundamento no domínio da enologia;

Estudo de mercado; Determinação da base de dados; Web Scraping/aquisição e preparação dos dados; Análise exploratória dos dados; Escolha das bibliotecas python; Abertura de repositório GitHub; Descrição da metodologia analítica e definição dos cálculos para verificação de acurácia; Consolidação de entrega. Até 30/10/2023: Desenvolvimento inicial do modelo; Parametrização; Treinamento do modelo; Verificação de acurácia com os métodos mencionados; Descrições dos resultados preliminares; Apresentação de modelo inicial; Descrição do modelo de negócio; Esboço de storytelling; Consolidação de entrega. Até 13/11/2023: Hiperparametrizações; Ajustes pró acurácia; Relatório técnico; GitHub com modelo, read.me, slides e demais arquivos do projeto; Resultados finais;

Consolidação de entrega.

Storytelling (.ppt) com apresentação em vídeo (YouTube);

7. CONSIDERAÇÕES FINAIS

REFERÊNCIAS

DE ALMEIDA, Maria João. O Vinho na Ponta da Língua. Leya, 2020.

PHILLIPS, Rod. Uma breve história do vinho. Editora Record, 2020.

PINTO DOS SANTOS, Norberto. **Novos e velhos territórios nos lazeres contemporâneos**. 2010.

GLOSSÁRIO