COMPONENTE CURRICULAR:	PROJETO APLICADO II - 2024.2	
NOMES COMPLETOS DOS ALUNOS:	Andre Gustavo Monteiro Dos Santos F	– RA 10424359
	Fernando da Silva Cordeiro de Lima	– RA 10424794
	Gabriel Santos de Oliveira	– RA 10424642
	Raul Santos Lages	– RA 10424621

PROJETO APLICADO II

GRUPO: CAMINHOS DA ESTATÍSTICA

Sumário

Objetivo do estudo: Apresentação da Empresa	3
Objetivo do estudo: Problema de Pesquisa	3
Cronograma de Atividades e Responsabilidades	5
Apresentação dos metadados e Análise Exploratória de Dados	6
Bibliotecas	7
Tratamento da Base	8
Definição e Descrição das Bases Teóricas dos Métodos	8
Definição e Descrição da Acurácia	10
Grupo	12
Github do projeto	12
Referências bibliográficas e sites consultados	12

Objetivo do estudo: Apresentação da Empresa

A empresa escolhida é a multinacional de tecnologia estadunidense Amazon, com o estudo se restringindo aos dados extraídos do site da mesma.

A Amazon foi fundada em 1994 e começou como uma livraria online, sendo uma das pioneiras neste setor. Hoje, é uma das maiores empresas do mundo por valor de mercado, com um leque de atuação extremamente diverso, composto por livros (tanto físico quanto digital), varejo, eletrônicos (como a assistente virtual Alexa e o leitor de livros Kindle), entretenimento por streaming (Prime Video e Twitch), video-games (Luna Cloud Gaming), serviços de cloud (AWS), entre outros.

Por se tratar de uma das maiores empresas de tecnologia do mundo e ter um vasto ramo de atuação, seus negócios possuem quantidades massivas de dados e algoritmos complexos para que a experiência do consumidor seja otimizada da melhor maneira possível.

Objetivo do estudo: Problema de Pesquisa

A Amazon, inicialmente uma modesta livraria online, transformou-se em uma gigante do varejo global, muito graças a sua capacidade de conquistar a confiança dos consumidores. No final dos anos 90 e início dos anos 2000, quando a internet ainda era vista com desconfiança — exacerbada pela bolha da internet —, a Amazon é um dos exemplos de superação de adversidade, pois, mesmo perdendo 90% do seu valor de mercado no início dos anos 2000, conseguiu sobreviver e se tornar referência no ramo da tecnologia.

Um aspecto central do crescimento da Amazon foi sua aposta em expandir seus produtos ofertados, trazendo para si parceiros externos, ou seja, tornando-se um market place. Para isso, o sistema de reviews foi essencial, junto com o nome da Amazon, foi essencial para que houvesse confiança dos consumidores. Afinal, é mais provável que um cliente opte por um produto avaliado e testado por outros usuários ou um produto sem avaliações? Além do mais, essa estratégia permitiu que a Amazon não apenas vendesse produtos, mas também oferecesse recomendações relevantes com base nos gostos e nas experiências dos clientes. Isso foi essencial para fortalecer ainda mais a confiança do consumidor e criar um ciclo de feedback contínuo que ajudou a empresa a se adaptar rapidamente às demandas dos clientes.

Esse sistema de avaliações será o foco desta pesquisa, especialmente no contexto do mercado de materiais de escritório, um setor que passou por grandes transformações durante a pandemia da COVID-19. A necessidade de adaptação ao home office fez com que muitas empresas e trabalhadores buscassem novos produtos para estruturar seus espaços de trabalho em casa. Nesse cenário, o papel das avaliações de produtos tornou-se ainda mais crucial, pois os consumidores, baseiam-se nas opiniões de outros usuários para escolher os melhores itens ou na sugestão dos departamentos de TI da empresa, que podem ter se baseado em opiniões dos usuários também. Produtos como cadeiras ergonômicas, mesas, equipamentos eletrônicos e material de escritório ganharam destaque nas avaliações, e a Amazon teve que ajustar sua oferta para atender a essa nova demanda – algo que ela conseguiu, tendo em vista que seu lucro subiu 220% com o crescimento das compras online durante a pandemia.

Portanto, o foco deste estudo está em analisar, por meio de modelos e medidas, como a Amazon, através de seu sistema de reviews, contribuiu para a adaptação de consumidores ao home office e como as avaliações de materiais de escritório ajudaram a guiar as decisões de compra durante esse período, além do impacto que a pandemia pode ter tido na procura por materiais de escritório. A pesquisa se divide nas seguintes etapas:

Coleta e pré-processamento de dados: utilizamos a base de dados "Amazon Reviews 2023", com foco em produtos do segmento de materiais de escritório ("Office_Products").

Desenvolvimento de um modelo de NLP: para classificar o sentimento das avaliações como positivo, negativo ou neutro, utilizamos modelos e/ou medidas como SVM, BERT, TF-IDF, entre outros.

Validação e teste: avaliar o desempenho do modelo em termos de precisão de recall usando um conjunto de dados de validação e ajustes conforme o necessário.

Implementação de um painel de controle: criar uma visualização que mostre a distribuição de sentimentos ao longo do tempo e oferecer insights sobre as mudanças nas preferências dos consumidores de materiais de escritório durante a pandemia.

Ao longo do estudo, será investigado como essas avaliações ajudaram os consumidores a encontrar produtos de qualidade e a superar as dificuldades impostas pela rápida transição para o home office, com foco no impacto desse sistema de reviews para a confiança e satisfação dos clientes no setor de materiais de escritório, utilizando metodologias escaláveis para estudo de outros tipos de produtos, o que pode ajudar

varejistas a entenderem o impacto de um bom sistema de avaliações, para que possam aprimorar ou criar o seu próprio, além de ajudar os comerciantes a priorizarem produtos, melhorar atendimento, e também guiar melhor o consumidor.

Cronograma de Atividades e Responsabilidades

Semana 1-2: Coleta e Pré-processamento dos Dados

- Coletar o dataset “Amazon Reviews 2023” para “Office_Products”.
- Explorar os dados para entender a estrutura e o conteúdo.
- Identificar e remover dados duplicados e irrelevantes.
- Realizar limpeza dos dados (remoção de stop words, pontuação, etc.).
- Realizar tokenização e normalização dos textos.
- Dividir os dados em conjuntos de treino, validação e teste.

Semana 3-4: Desenvolvimento do Modelo

- Selecionar e configurar a ferramenta de NLP (SVM, BERT, TF-IDF).
- Treinar um modelo inicial com um subconjunto dos dados.
- Ajustar hiperparâmetros do modelo para melhorar o desempenho.
- Implementar técnicas de aumento de dados, se necessário.

Semana 5-6: Validação e Teste

- Avaliar o modelo usando o conjunto de validação.
- Calcular métricas de desempenho (precisão, recall, F1-score).
- Ajustar o modelo com base nos resultados de validação.
- Realizar testes finais com o conjunto de teste.

Semana 7-8: Implementação e Apresentação

- Desenvolver um painel de controle para visualização dos resultados.
- Integrar o modelo de análise de sentimentos ao painel.

- Testar o painel com novos dados de avaliações.
- Preparar a apresentação final do projeto, destacando os resultados e insights obtidos.

Apresentação dos metadados e Análise Exploratória de Dados

Os dados foram obtidos através do link https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023 e serão utilizadas as bases em “Office_Products”. Abaixo, as colunas que serão analisadas, seus tipos e uma explicação de cada uma delas, conforme a fonte dos dados:

User Reviews

Campo	Tipo	Definição
rating	float	Nota do produto (de 1.0 a 5.0).
title	str	Título da avaliação do usuário.
text	str	Texto da avaliação do usuário.
images	list	Imagens postadas pelos usuários após recebimento do produto. Cada imagem tem tamanhos diferentes (pequeno, médio, grande), representadas por small_image_url, medium_image_url e large_image_url, respectivamente.
asin	str	ID do produto.
parent_asin	str	Parent ID do produto. Observação: variações de um produto geralmente pertencem a um mesmo Parent ID. O campo “asin” em datasets mais antigos na verdade é o Parent ID, então o Parent ID que será utilizado para encontrar o metadado.
user_id	str	ID do consumidor que avaliou o produto.
timestamp	int	Data da avaliação (horário Unix)
verified_purchas	bool	Verificação de compra do usuário.
helpful_vote	int	Número de vezes que usuários classificaram a avaliação como "Útil".

Tabela 1: campo, tipo e definição de User Reviews.

Item Metadata

Campo	Tipo	Definição
main_category	str	A categoria mais relevante em que o produto se encontra.
title	str	O título que identifica o produto.
average_rating	float	Classificação do produto com base nas avaliações dos clientes.
rating_number	int	Quantidade de pessoas que deixaram uma avaliação para o produto.
features	list	Características principais do produto destacadas em formato de bullet-points.
description	list	Informações detalhadas sobre as especificações do produto.
price	float	Valor do produto em dólares no momento da coleta.
images	list	Fotos do produto em diferentes tamanhos (miniatura, grande, alta resolução), com o campo "variante" indicando a posição da imagem.
videos	list	Título e URL dos vídeos relacionados ao produto.
store	str	Identificação do vendedor do produto.
categories	list	Classificação do produto dentro de um sistema de categorias.
details	dict	Informações sobre materiais, marca, tamanhos e outras especificações do produto.
parent_asin	str	Identificação do produto pai a que o item pertence.
bought_together	list	Sugestões de pacotes recomendados pelos sites para compra junto com o produto.

Tabela 2: campo, tipo e definição de Item Metadata.

Análise Exploratória de Dados

A análise exploratória é fundamental para entender a estrutura e as características dos dados. Nesta seção, utilizaremos gráficos e estatísticas descritivas para compreender melhor os dados para que o modelo seja aplicado posteriormente e seja possível.

Gráficos:

- Histogramas para as distribuições das notas.
- Gráficos de dispersão para correlação entre notas e sentimentos.
- Boxplots para identificar outliers nas notas.

Estatísticas Descritivas:

- Média, mediana, e moda das notas.
- Frequência de categorias de sentimentos em relação às notas.

Análise:

- Identificação de padrões e tendências nos dados.

Bibliotecas

Vamos utilizar as seguintes bibliotecas e suas funções específicas:

nltk: Utilizada para o processamento de linguagem natural, incluindo a tokenização, stemming, lematização e remoção de stopwords.

plotly : Excelente para criar gráficos interativos e visualizações dinâmicas.

pandas: Essencial para a manipulação e análise de dados, facilitando a limpeza, transformação e exploração de datasets.

seaborn: Complementa o matplotlib, fornecendo gráficos estatísticos e estéticos aprimorados.

watermark: Usada para adicionar informações de rodapé como versão de biblioteca e data nos notebooks Jupyter.

numpy: Oferece suporte para arrays multidimensionais e funções matemáticas de alto desempenho.

re: Biblioteca para manipulação e operação de expressões regulares, essencial para encontrar padrões e substituir textos.

matplotlib: Ferramenta para criação de gráficos estáticos, animados e interativos, frequentemente utilizada em conjunto com pandas e numpy.

Com estas bibliotecas, abordaremos de forma eficiente todas as etapas necessárias para nossa análise.

Tratamento da Base

Limpeza de Dados:

- Remoção de dados duplicados.
- Tratamento de valores ausentes (imputação ou exclusão).
- Remoção de comentários não relevantes ou spam.

Transformação de Dados:

- Conversão de notas em categorias (positivo, neutro, negativo).
- Pré-processamento de texto (remoção de stopwords, stemming ou lemmatization).
- Criação de variáveis dummy, se necessário.

Divisão dos Dados: Separação dos dados em conjuntos de treino e teste.

Definição e Descrição das Bases Teóricas dos Métodos

O estudo irá abordar a problemática por meio de três modelos/técnicas: Support Vector Machines (SVM), Bidirectional Encoder Representations from Transformers (BERT) e Term Frequency-Inverse Document Frequency (TF-IDF).

Support Vector Machines (SVM)

O SVM (Support Vector Machine ou Máquina de Vetores de Suporte) é um algoritmo de aprendizado de máquina voltado para a classificação, que procura identificar o hiperplano ideal que maximiza a distância entre classes distintas. Esse hiperplano divide os dados em um espaço de n dimensões, enquanto as linhas adjacentes a ele, conhecidas como vetores de suporte, são os pontos mais próximos que determinam essa distância. Ao maximizar essa margem, o SVM consegue generalizar de forma mais eficaz, proporcionando previsões de classificação mais precisas em novos conjuntos de dados.

Bidirectional Encoder Representations from Transformers (BERT)

O BERT (Bidirectional Encoder Representations from Transformers) é um algoritmo de aprendizado profundo que serve como um “tradutor” da linguagem dos seres humanos para a máquina. Ele utiliza a arquitetura Transformer, um mecanismo de atenção que aprende as relações contextuais entre palavras (ou subpalavras) em um texto. Ao contrário de modelos direcionais que leem o texto sequencialmente, o encoder do Transformer lê toda a sequência de palavras de uma vez, permitindo que o modelo aprenda o contexto de uma palavra com base em seu entorno, tanto à esquerda quanto à direita. Para treinar o modelo de linguagem, o BERT emprega duas estratégias de treinamento: a previsão de tokens mascarados e a previsão de próxima frase, superando assim as limitações de modelos que apenas preveem a próxima palavra em uma sequência.

Term Frequency-Inverse Document Frequency (TF-IDF)

O TF-IDF (Term Frequency-Inverse Document Frequency) é uma medida utilizada para quantificar a importância ou relevância de textos em um documento em relação a uma coleção de documentos.

O TF significa a frequência dos termos, seja ela a quantidade bruta que esses termos aparecem no documento, ou pela frequência com base no tamanho do documento ou mesmo de forma booleana, considerando 1 para o termo que aparece e 0 se não aparece.

O IDF mostra a raridade dos termos em relação à coleção de documentos. Ele pega o total de documentos na coleção e analisa quantos possuem os termos procurados, fazendo uma divisão de forma logarítmica do número total de documentos na coleção pelo número de documentos que possuem os termos.

O resultado final, TF-IDF, é a multiplicação de ambos os termos, portanto, quanto mais próximo de 0 for o resultado, menos relevante será o termo.

Definição e Descrição da Acurácia

A acurácia é uma medida crucial na avaliação de modelos de análise de sentimentos, especialmente quando se busca entender a relação entre as avaliações numéricas (notas de 0 a 5) e os comentários textuais dos usuários. Neste contexto, a acurácia pode ser definida como a proporção de previsões corretas sobre o total de previsões feitas. Para avaliar a correspondência entre a nota atribuída e a análise de sentimento dos comentários, consideramos os seguintes pontos:

Categorização das Notas

As notas podem ser categorizadas como positivas (4 e 5), neutras (3) e negativas (0 a 2). Essa categorização nos permite comparar a polaridade dos comentários com a nota fornecida.

Análise de Sentimentos

Utilizamos um modelo de análise de sentimentos para classificar os comentários em categorias semelhantes: positivo, neutro ou negativo. O modelo é treinado com dados rotulados, onde os sentimentos são previamente definidos.

Métricas de Avaliação

Para avaliar a eficácia do modelo, utilizamos a matriz de confusão, que nos permite visualizar o desempenho do modelo em termos de previsões corretas e incorretas. As métricas que analisamos incluem:

True Positives (TP): Comentários positivos que correspondem a notas de 4 ou 5.

True Negatives (TN): Comentários negativos que correspondem a notas de 0 a 2.

False Positives (FP): Comentários positivos que, na realidade, correspondem a notas de 0 a 2.

False Negatives (FN): Comentários negativos que correspondem a notas de 4 ou 5.

Acurácia do Modelo

A acurácia do modelo é calculada com a fórmula:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

Imagem 1: fórmula da acurácia.

Essa métrica nos proporciona uma visão geral de quão bem o modelo está correlacionando as notas com os sentimentos expressos nos comentários.

Análise de Resultados

Além da acurácia, é essencial considerar outras métricas, como precisão e recall, para entender a performance do modelo de maneira mais holística. A precisão nos informa sobre a proporção de previsões positivas que foram corretas, enquanto o recall nos mostra a proporção de sentimentos positivos que foram corretamente identificados. A média harmônica entre essas duas métricas, o F1-score, também pode ser uma indicação valiosa da eficácia do modelo.

Ao utilizar uma combinação de métricas de acurácia, precisão, recall e F1-score, podemos validar se os usuários se expressaram corretamente em seus comentários ou se, por outro lado, a interpretação do sentimento pelo modelo falhou em refletir a avaliação real. Essa abordagem fornece uma visão mais clara do comportamento dos consumidores e pode informar melhorias em produtos e serviços.

Grupo

O grupo é formado por:

Andre Gustavo Monteiro Dos Santos F – RA 10424359

Fernando da Silva Cordeiro de Lima – RA 10424794

Gabriel Santos de Oliveira – RA 10424642

Raul Santos Lages – RA 10424621

Github do projeto

https://github.com/rourp/caminhos-da-estatistica

Referências bibliográficas e sites consultados

ALQAHTANI, Arwa S. M. **PRODUCT SENTIMENT ANALYSIS FOR AMAZON REVIEWS**. 2021.

Amazon's Profits Triple – https://www.nytimes.com/2021/04/29/technology/amazons-profits-triple.html – Acesso em 20/09/2024.

BERT (language model) – https://en.wikipedia.org/wiki/BERT_(language_model) – Acesso em 25/09/2024.

BERT Explained: State-of-the-art language model for NLP – https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 – Acesso em 25/09/2024

DANG, Nhan Cach; MORENO-GARCÍA, María N.; PRIETA, Fernando de La. **Sentiment Analysis Based on Deep Learning: A Comparative Study**. 2020.

DEVLIN, Jacob et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

FLETCHER, Tristan. **Support Vector Machines Explained**. 2008.

History of Amazon – https://en.wikipedia.org/wiki/History_of_Amazon – Acesso em 20/09/2024.

List of Amazon products and services – https://en.wikipedia.org/wiki/List_of_Amazon_products_and_services – Acesso em 20/09/2024.

Support Vector Machine – https://www.ibm.com/topics/support-vector-machine – Acesso em 25/09/2024

Understanding TF-IDF – https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/ – Acesso em 25/09/2024

Vinte anos depois da bolha da internet, as sobreviventes viraram trilionárias – https://exame.com/tecnologia/vinte-anos-depois-da-bolha-da-internet-as-sobreviventes-viraram-trilionarias/ – Acesso em 20/09/2024

Why is Amazon so successful and how did it get here? – https://www.amazonppc.co.uk/post/why-is-amazon-so-successful-and-how-did-it-get-here – Acesso em 20/09/2024