MODELO PARA A ENTREGA DAS ATIVIDADES

COMPONENTE CURRICULAR:	PROJETO APLICADO II
NOME COMPLETO DO ALUNO:	Clayton dos Santos Lira Lorena Vaz Cord Tiago Clemente Rodrigues Lucas Vaz de Castro Oliveira
RA:	10416054 10424700 10423746 10424623

Atenção: Toda atividade deverá ser feita com fonte Arial, tamanho 11, espaço de 1,5 entre as linhas e alinhamento justificado entre as margens.

UNIVERSIDADE PRESBITERIANA MACKENZIE

Tecnologia em ciência de dados

Análise de Sentimento em Comentários de E-commerce Usando Processamento de Linguagem Natural e Aprendizado de Máquina

São Paulo

2024

Introdução

O comércio eletrônico, ou *e-commerce*, tem se consolidado como uma das formas mais relevantes e dinâmicas de comércio em todo o mundo, promovendo o crescimento da economia digital e transformando os padrões de consumo. O Brasil é um dos maiores mercados de e-commerce da América Latina, e entender as opiniões dos clientes é crucial para manter a competitividade e oferecer experiências excepcionais.

Com o uso de técnicas de Processamento de Linguagem Natural (PLN), torna-se possível automatizar a análise de grandes volumes de dados textuais, avaliando as opiniões dos clientes permitindo respostas rápidas a feedbacks negativos.

Neste contexto, propomos avaliar os comentários de compras realizadas na Olist, no período de 2016 a 2018. O objetivo é identificar áreas de melhoria na experiência dos clientes, promovendo uma compreensão mais profunda de suas emoções e contribuindo para a melhoria contínua dos serviços e produtos oferecidos pela empresa.

Cronograma de trabalho

Atividade	Datas estimadas
Pré-processamento dos dados e análise exploratória dos dados	16/09/2024 - 20/09/2024
Treinamento do modelo	20/09/2024 - 27/09/2024
Validação e ajustes	28/09/2024 - 04/10/2024
Análise dos resultados	04/10/2024 - 10/10/2024
Elaboração de relatórios	11/10/2024 - 18/10/2024
Elaboração do trabalho escrito	01/09/2024 - 20/11/2024
Elaboração dos slides e apresentação final	10/11/2024 - 20/11/2024

Objetivo geral do estudo

Realizar, por meio da linguagem python, uma análise de sentimento em comentários dos clientes a partir das avaliações e comentários feitos pelos clientes em uma base com dados de e-*commerce*, utilizando técnicas de processamento de linguagem natural e aprendizado de máquina para identificar padrões de opinião e gerar insights que possam apoiar a tomada de decisões estratégicas.

Objetivos específicos

1. Realizar o pré-processamento linguístico dos textos, por meio de processos de tokenização, remoção de stopwords e lematização para transformá-los em inputs adequados para modelagem;
2. Utilizar as técnicas POS-tagging e a representação vetorial TF-IDF para converter os textos em representações numéricas para uso no modelo proposto;
3. Treinar e avaliar um modelo de aprendizado de máquina, utilizando algoritmo Random Forest com o intuito de classificar sentimentos e avaliar seu desempenho utilizando as métricas de acurácia, precisão, revocação e F1-score.

Empresa

A Olist é uma empresa que oferece um conjunto completo de soluções voltadas para negócios. Entre os serviços oferecidos, destaca-se a Olist Store, uma plataforma de vendas online que opera em diversas lojas e marketplaces populares no Brasil, como Amazon, Shopee, Magalu, Americanas, entre outros. Através da Olist, empresas têm a oportunidade de expor seus produtos nos sites de grandes redes de varejo, contando com suporte em áreas como gestão, logística e atendimento ao cliente. A Olist não comercializa produtos próprios, mas possibilita que lojistas e fabricantes utilizem sua plataforma para impulsionar suas vendas. Isso é possível porque a empresa possui grande relevância nesses marketplaces, o que resulta em maior visibilidade para os parceiros. Dessa forma, a Olist recebe uma comissão sobre as vendas, enquanto seus parceiros conseguem ampliar suas operações e aumentar o volume de vendas mensais. Para mais informações, é possível acessar o site oficial da Olist.

Referências de aquisição do dataset

O conjunto de dados utilizado neste projeto foi disponibilizado pela empresa proprietária dos dados por meio da plataforma Kaggle, renomada e conhecida como uma comunidade online voltada para competições e troca de conhecimentos entre profissionais e entusiastas de ciência de dados. Ele pode ser acessado através do link https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Esse dataset foi publicado sob a licença pública Creative Commons CC BY-NC-SA 4.0 (Atribuição-NãoComercial-CompartilhaIgual 4.0 Internacional), o que permite a reprodução e o compartilhamento dos dados originais ou modificados, desde que seja para fins não comerciais.

Descrição do dataset e metadados

De acordo com o proprietário dos dados, o dataset fornecido contém informações detalhadas sobre clientes, sua localização, vendedores, valores de frete e produtos, além das avaliações e comentários de clientes a respeito das compras realizadas através da plataforma. O conjunto abrange aproximadamente 100.000 pedidos realizados entre 2016 e 2018 em diversos marketplaces no Brasil. Embora os dados sejam reais, a identidade dos clientes e parceiros comerciais foi preservada, com os nomes substituídos por referências da série de televisão Game of Thrones, garantindo o anonimato.

O conjunto de dados é composto por nove tabelas no formato CSV. Contudo, este estudo se concentra especificamente na quinta tabela, denominada "olist_order_reviews_dataset.csv". A seguir, são apresentadas a descrição dos atributos desta tabela.

1. olist_order_reviews_dataset.csv

Descritivo: Dados com avaliações e comentários feitos pelos clientes.

Linhas: 99.224

Atributos:

- review_id (string)
- order_id (string)
- review_score (integer)
- review_comment_title (string)
- review_comment_message (string)
- review_creation_date (datetime)
- review_answer_timestamp (datetime)

Análise exploratória

Este estudo será realizado com base no conjunto de dados "olist_order_reviews_dataset" e seus conteúdos, utilizando a linguagem Python. A análise será conduzida no Google Colab, e os resultados serão carregados em nosso repositório no GitHub. Todo o estudo, incluindo explicações, estará documentado nos próprios códigos e notas do Colab.

Na primeira etapa deste trabalho, focamos no tratamento dos dados e das bases (os códigos e detalhes podem ser encontrados no GitHub). Durante o processo de tratamento, realizamos as seguintes atividades:

- Verificamos as colunas e linhas dos dataframes utilizando os comandos head e shape.
- Combinamos os valores do título dos comentários com o texto correspondente, enriquecendo assim a análise que será realizada.
- Removemos valores ausentes da combinação de título e texto dos comentários, a fim de evitar viés em nossa análise.

Com os dados tratados, avançaremos para a análise exploratória propriamente dita, onde buscaremos extrair insights significativos. Algumas das perguntas que guiaram nossa investigação incluem:

- Qual é a distribuição do ‘review_score’?
- Qual é a distribuição de sentimentos ‘positivo’ e ‘negativo’?

Utilizaremos gráficos para ilustrar essas informações. A partir dos dados tratados, começaremos a examinar as informações disponíveis em nosso conjunto de dados.

Tratamento da base de dados

Observamos um desequilíbrio entre as classes de sentimentos: as avaliações ‘positivas’ totalizam 26.530, enquanto as ‘negativas’ somam 10.890. Para evitar viés na análise, utilizaremos o método de reamostragem (resampling) para subamostrar as avaliações positivas, resultando em 10.890 avaliações para cada classe.

Realizamos o tratamento dos dados textuais combinados, removendo a capitalização das letras, emojis, acentuação, URLs e espaços em branco indesejados nos comentários. Em seguida, executamos as etapas de tokenização, remoção de stopwords, lematização, POS-tagging e representação vetorial utilizando o modelo de Term Frequency — Inverse Document Frequency (TF-IDF).

Após esse processamento, separamos o conjunto de dados em variáveis dependentes e independentes e, logo depois, dividimos os dados em conjuntos de treinamento e teste para a aplicação de técnicas de aprendizado de máquina (ML).

Optamos por um modelo de classificação de floresta aleatória, que é um algoritmo amplamente utilizado em tarefas de classificação devido à sua capacidade de lidar com dados complexos e à resistência ao overfitting. Por fim, avaliamos o desempenho do modelo treinado por meio de métricas como precisão, recall, F1-Score e acurácia.

Observando os resultados obtidos podemos ver que o modelo apresenta um desempenho muito bom, com alta precisão, recall e F1-Score para ambas as classes. A acurácia geral de 92% e os altos valores de precisão e recall sugerem que o modelo é eficaz na classificação de avaliações como "positivo" ou "negativo". Além disso, as médias (macro e weighted) reforçam que o modelo é consistente em seu desempenho, independentemente do suporte de cada classe. Isso torna o modelo bastante confiável para o propósito de análise de sentimentos.

Conclusão

A análise de sentimentos tem se tornado uma ferramenta essencial para empresas que desejam entender melhor a percepção de seus clientes, e a Olist, uma plataforma de e-commerce que conecta vendedores a grandes marketplaces, é um exemplo claro dessa aplicação.

A Olist não apenas facilita a venda de produtos através de sua plataforma, mas também coleta uma vasta quantidade de dados sobre as experiências dos consumidores. Com a crescente concorrência no setor de e-commerce, compreender as emoções e opiniões dos clientes se tornou crucial para manter a competitividade. A análise de sentimentos permite que a Olist identifique rapidamente áreas que necessitam de melhorias, respondendo de forma proativa a feedbacks negativos e aprimorando a experiência do usuário.

Utilizando técnicas avançadas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, a Olist pode analisar comentários e avaliações feitas por clientes. Por exemplo, ao aplicar um modelo de floresta aleatória para classificar sentimentos em avaliações, a empresa consegue não apenas determinar se uma opinião é positiva ou negativa, mas também extrair insights valiosos sobre o que os clientes realmente pensam sobre seus produtos e serviços.

O impacto da análise de sentimentos na Olist pode ser significativo. Com uma acurácia geral do modelo em torno de 92%, a empresa pode confiar nas classificações geradas para tomar decisões informadas sobre melhorias em seus serviços. Isso não apenas melhora a satisfação do cliente, mas também fortalece a reputação da marca no mercado. Ao entender melhor as emoções dos consumidores, a Olist pode não apenas melhorar sua oferta, mas também se posicionar como uma líder no competitivo cenário do e-commerce brasileiro.

Repositório do Github

Endereço para o repositório: https://github.com/lucasvazcastro/Projeto_Aplicado_II