

Análise de Comentários em E-commerces

Jailson da Costa Dias (jcd2) e Ramom Pereira dos Santos Silva (rpss)

1. Introdução

Com a popularização das vendas online, tanto em e-commerces como a Amazon, quanto nas redes sociais como Instagram, Facebook e WhatsApp, surge uma quantidade imensa de avaliações de produtos que muitas vezes são difíceis de identificar o que é dito nessas avaliações, pois a quantidade é grande e existem muitas avaliações que não são boas para extrair alguma informação sobre o produto. Assim surge um grande problema que é identificar quais os problemas ocorridos durante o processo de compra ou até mesmo depois da compra, para assim conseguir dar uma melhor experiência de compra para o cliente.

2. Coleta de Dados

Para fazer esse projeto vai ser utilizado uma base de dados do kaggle que pode ser encontrada no seguinte link <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>, esta é uma base sobre avaliações e comentários de produtos vendidos online. Essa base tem informações do ID da avaliação, ID do produto avaliado, texto da avaliação em inglês, classe do produto entre outras coisas, no total tem 10 features.

3. Hipóteses

1. Com a idade da pessoa e a avaliação que ela deu ao produto, podemos inferir qual é a classe do produto;
2. Pessoas jovens entre 18 e 30 anos preferem vestidos;
3. Produtos mais bem avaliados são mais vendidos;
4. Produtos bem avaliados têm maiores quantidades de comentários.

4. Pré-processamento da base

Nesta base vamos executar um algoritmo de processamento de linguagem natural para identificar algumas características no comentário, que são elas, identificar se o comentário é relevante, ou seja, se ele está falando alguma coisa sobre o produto ou sobre a loja, e se ele está falando positivamente ou negativamente. O objetivo dessa etapa é retirar todos os comentários de apenas uma palavra ou comentários que não estão falando do produto ou da loja e saber sobre a experiência do usuário no processo de compra e no pós-compra.

Além disso, vamos fazer uma análise da idade das pessoas e categorizar em três categorias, pessoas até 25 anos (jovens), pessoas entre 25 e 60 anos (adulto) e acima de 60 anos (terceira idade). Vamos também, dividir a avaliação que está numérica de 1 a 5 em categórica, onde 1 e 2 é muito ruim, 3 e 4 razoável e 5 é boa.

5. Análise de Dados

A partir desta nova base de dados vai ser feito uma visualização separada por tipo de comentário organizando em, comentários sobre a loja, sobre o produto e que não está falando especificamente de nenhum dos dois. Depois dessa visualização vai ter outra visualização para mostrar se o comentário está falando positivamente ou negativamente, neste caso vai ser analisado apenas as categorias de loja e produto. No caso da categoria que não está falando especificamente de nenhum dos dois, não vai ser feita essa segunda análise.

Faremos visualizações de todas as features da base de dados, buscaremos também relações entre as features para analisar, por exemplo, produtos mais comprados em conjunto e os produtos mais comprados por cada categoria de pessoa. Com todas essas visualizações vamos verificar se as hipóteses propostas estão corretas ou erradas e criar novas hipóteses sobre o conjunto de dados e analisá-las.

Abaixo pode ser visto como vai ser uma visualização, a visualização dos comentários, que nela vai ter 2 categorias e para cada categoria vai ter duas sub categorias de comentários.

Visualização de comentários:

1. Comentários falando da loja
 - a. Positivamente
 - b. Negativamente
2. Comentários falando do produto
 - a. Positivamente
 - b. Negativamente