Quais são os fatores que mais aborrecem os clientes em suas compras online?

Murilo Menezes (Insper, São Paulo, Brasil).
Repositório: https://github.com/murilomenezes1/NLP-AI

1 Introdução

Uma dos principais problemas enfrentados por varejistas é avaliar seus produtos e entender o que pode ser melhorado para que a conversão aumente. Para isso, é necessário uma quantidade relevante de dados, o que nem sempre é fácil de conseguir em alguns casos. Para o e-commerce, o cenário tende a ser mais favorável, uma vez que o processo de avaliação, para o cliente, é mais intuitivo e de fácil acesso, melhorando o processo para que um lojista tenha acesso aos dados relevantes. Entretanto, o principal sistema para captação das avaliações é o de "estrelas", que podem ou não ser acompanhadas de um comentário, e estes são não-estruturados, o que dificulta a padronização de um método para avaliar o sentimento por trás de cada comentário. Com isto em mente, este texto descreve uma abordagem de solução para o problema, que parte de algumas premissas. Primeiramente, assume-se que avaliações de 2 estrelas ou menos são necessariamente negativas. Além disso, avaliações de 3 ou mais estrelas são necessariamente neutras ou boas. Considera-se também que os comentários contém informações relevantes para o problema.

2 Metodologia

Para a implementação da solução, iniciou-se lendo o dataset e limpando os dados, removendo as linhas com valores "NaN", uma vez que estamos interessados no conteúdo em texto das colunas, e dados que fogem deste quadro não acrescentam na análise. Em seguida, é feito um recorte do dataset, contendo apenas es avaliações de 2 estrelas ou menos, a fim de focarmos nas avaliações necessariamente negativas. Para a análise, utilizou-se um "CountVectorizer", que transforma os dados em vetores, separando cada palavra dentro de um comentário. Em seguida, o processo de "fit-transform" cria um vocabulário de tokens únicos para os dados de entrada e faz a transformação dos dados para uma representação numérica que contém a frequência de um determinado token nos dados analisados. A fim de enriquecer a análise, também foi feita uma nova iteração, implementando a análise de bigramas e trigramas, a fim de adicionarmos um componente de contexto para a solução, o que ajuda a entender melhor os fatores que aborrecem os clientes, além de permitir que seja feita uma validação dos comentários que são relevantes.

3 Resultados

A partir da primeira análise, obteve-se um gráfico que indica as 60 palavras mais frequentes nos dados, como mostra a Figura 1.

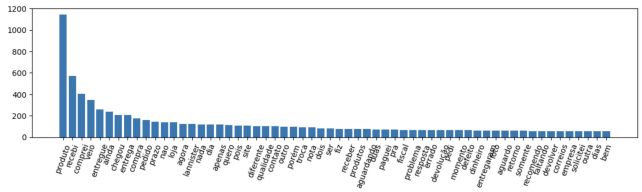

Figure 1: Frequência de tokens únicos nas avaliações de 2 estrelas ou menos

Nota-se que, para o recorte de 2 estrelas ou menos, dos 5 tokens mais frequentes, 3 estão relacionados à entrega do produto ("Recebi", "Veio" e "Entregue"). Além disso, ao analisarmos as demais palavras no gráfico, é possivel identificar algumas outras que remetem à entrega, como "prazo" e "dia", mas a falta de um contexto para as palavras limita a profundidade da análise.
Com a terceira análise, obteve-se um novo gráfico, que identifica os conjuntos de palavras com maior frequência no dataset, através do argumento de "ngrams".

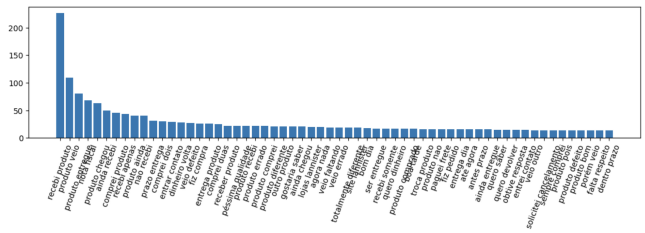

Figure 2: Frequência de bigramas nas avaliações de 2 estrelas ou menos (ngrams = 2)

Com o componente de contexto implementado, é confirmado que o principal problema enfrentado pelos consumidores está relacionado com o processo de entrega, mas ainda não é claro se está relacionado à demora, qualidade, etc.
Conforme o coeficiente de "ngrams" é aumentado, maior é a capacidade de entendimento do contexto das avaliações e, quando avaliamos os trigramas, fica nítido que a insatisfação dos clientes está relacionada principalmente à demora na entrega, mas nota-se que os problemas ocorrem por todo o processo de logística, uma vez que muitos clientes reportam que receberam o produto errado, ou que este veio com defeito, como mostra a figura 3.

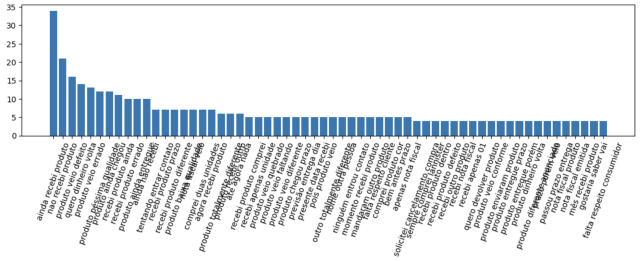

Figure 3: Frequência de trigramas nas avaliações de 2 estrelas ou menos (ngrams = 3)