Avaliação Intermediária Natural Language Processing

Matheus Freitas Sant'Ana (Insper, São Paulo, Brasil)

1 Introdução

Em compras online, diversos problemas podem surgir. Uma forma de obter um feedback sobre a entrega é por meio de avaliações, que podem englobar diversos elementos. Nesse projeto, será buscado entender quais são os principais motivos do porquê consumidores não gostaram de uma entrega. Para tal, pressupõe-se que para descobrir esses principais fatores será necessário descobrir quais são as palavras mais relevantes em críticas negativas.

2 Análise dos dados

O dataframe que foi analisado nesse projeto disponibilizava diversas colunas, portanto foi necessário primeiramente um processo de limpeza de dados, filtrando apenas aqueles que foram considerados mais relevantes. As colunas de IDs do pedido e da avaliação, bem como a data e a hora em que ela foi feita, foram descartadas. Os dados remanescentes eram sobre o título de uma crítica, sua mensagem (descrição) e também uma nota que varia de 1 a 5. Verificou-se que muitas vezes uma avaliação não possuía título, porém, levou-se em conta esses dados também, uma vez que pode funcionar como um resumo da avaliação, englobando palavras importantes (relevantes). Isto é, mesmo que não houvessem muitos títulos, eles são importantes na medida em que irão conter palavras que provavelmente também aparecerão na descrição (no caso em que os dois existem), e um maior o número de vezes que a palavra aparecer contribui para que ela seja mais relevante. Dessa forma, foi necessário dividir as avaliações em "positivas" e "negativas" de acordo com algum critério. No caso, o critério foi escolhido de acordo com os dados de scores de cada avaliação, estabelecendo um limite de que se a nota atribuída a uma entrega foi maior ou igual a 3, ela é uma avaliação positiva, e negativa caso contrário. Isso foi feito com base em uma análise empírica em que algumas amostras foram observadas, e decidiu-se que esse limite de 3 era ideal, visto que abaixo disso as avaliações costumavam ser expressiva e visivelmente negativas, com muitas reclamações. As notas três eram mais neutras, no entanto, julgou-se ser importante incorporá-las aos dados, pois dessa forma haveria uma perda menos considerável de dados, o que ajudaria a encontrar as palavras mais relevantes que é o principal objetivo.

3 Cálculo da relevância

A estratégia utilizada foi a de criar duas listas, uma para as todas as frases positivas e outra para as negativas (sendo que os dados englobados eram aqueles dos títulos e descrições das avaliações). Apesar do interesse da pesquisa ser o de aborrecimento dos consumidores, julgou-se importante também encapsular as críticas positivas, para entender também quais qualidades faltaram para uma avaliação receber uma nota boa. Além disso, também as positivas foram levadas em conta, pois julgou-se que seria uma boa maneira de filtrar palavras que

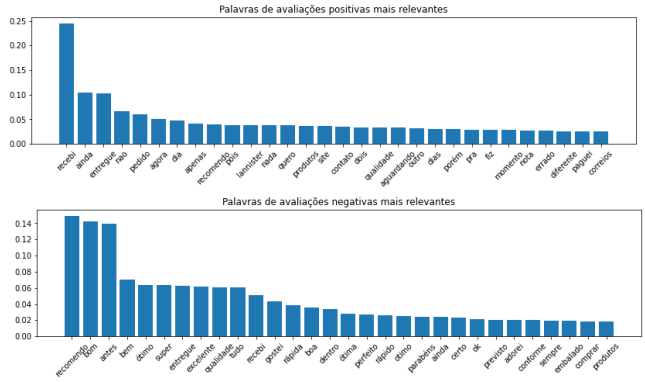

Figure 1: Palavras positivas e negativas mais relevantes

aparenemente são relevantes, mas no fundo não são (falsos positivos). Foi pressuposto que esse fato poderia ser observado ao notar quais palavras eram relevantes em ambos as listas. Para calcular a relevância, utilizou-se o CountVectorizer (individualmente para positivos e negativos) para calcular o TFIDF de cada palavra. Vale ressaltar também que uma lista de stopwords (palavras a serem ignoradas) foi utilizada, cuja fonte está no código que está referenciado no final desse documento. A partir disso, um gráfico foi gerado, com as palavras no entanto, percebeu-se muitas similaridades entre o gráfico positivo, sustentando a tese de que poderiam aparecer falsos positivos nas palavras mais relevantes. Isso pode acontecer também devido ao fato de que muitas frases negativas podem conter advérbios que invertem completamente o sentido da frase, mas utiliza palavras comuns a ambas positivas e negativas e por isso muitas palavras apareceram nos dois gráficos. Com o objetivo de contornar esse problema, algumas stopwords "extras" foram adicionadas manualmente à lista, como "prazo", "entrega", "produto", a fim de refinar a classificação. Com isso, obteve-se o resultado que pode ser visto na figura 1 acima, em que as palavras mais relevantes aparecem.

4 Conclusão

Com a figura obtida, palavras como "aguardando" e "dias" explicitam que uma das maiores causas das notas negativas é o atraso em relação ao prazo estipulado, causando irritação nos consumidores. "Errado" e "diferente" levantam outra possibilidade que é pode ter vindo algo que não o prometido. "Correios", "contato" podem levantar a hipótese de que problemas de comunicação e logística também causam muita irritação. Já nas palavras positivas observa-se a presença de muitos adjetivos de elogios como "ótimo", "perfeito", "excelente". Com isso, pode-se afirmar que a busca por palavras por relevância mostrou-se eficiente, mas pode ser melhorado ao utilizar Stemming para remover sufixos e evitar redundaância como "dia" e "dias". Segue link do repositório: https://github.com/MatFreitas/AI_NLP.git