Análise de comentários em E-commerces

Jailson Dias e Ramom Pereira

Universidade Federal de Pernambuco

Resumo Este documento tem por objetivo analisar o comportamento de pessoas que compram roupas femininas em e-commerces, através dos reviews feitos de produtos e lojas. Foram feitas análises para entender o que cada categoria de pessoa mais compra e satisfação de cada usuário com a compra.

1 Introdução

A partir do conjunto de dados contendo os reviews, foram feitas várias análises das características dessas compras, por exemplo produtos mais comprados por jovens, satisfação deles com essas compras, o perfil dos usuários. Por fim, foram utilizados os algoritmos, Árvore de Decisão, KNN e Naive Bayes, para predizer os produtos comprados por uma pessoa através de seu comentário.

2 Metodologia

2.1 Coleta de Dados

Os dados foram coletados do portal Kaggle, que possui diversos datasets dos mais variados temas, mas nesse projeto foi escolhido um dataset de reviews de compras de roupas femininas em e-commerces. O conjunto de dados pode ser encontrado em: https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews.
Dentre os dados desse dataset destacam-se as seguintes características:

(i) Idade da pessoa que está avaliando;
(ii) O review propriamente dito, em inglês;
(iii) Classe do produto avaliado;
(iv) Nota dada pelo usuário.

3 Pré-Processamento dos Dados

Como está lidando com texto, foi necessário executar processamento de linguagem natural, para classificar os comentários em categorias, por exemplo se é sobre a loja ou produto, como também para classificar em positivos e negativos. Para isso, foi o IBM Watson com 100 exemplos da base dados para treiná-lo. Foi

necessário também, alterar os tipos de alguns atributos, em sua maioria transformando em atributos categóricos para ter uma análise mais precisa. Atributos como, a nota do usuário tornaram-se categóricos, assim como, foram criados algumas colunas necessárias para a análise, como a categoria da pessoa, em relação a idade e se a nota dada é ruim, neutra ou boa. Além disso, foi feita uma comparação da nota dada com a o resultado do processamento de linguagem natural.

3.1 Análise Exploratória

A análise começou verificando o perfil em relação a idade dos usuários, nesse caso a maioria era adulta, entre 30 e 50 anos.

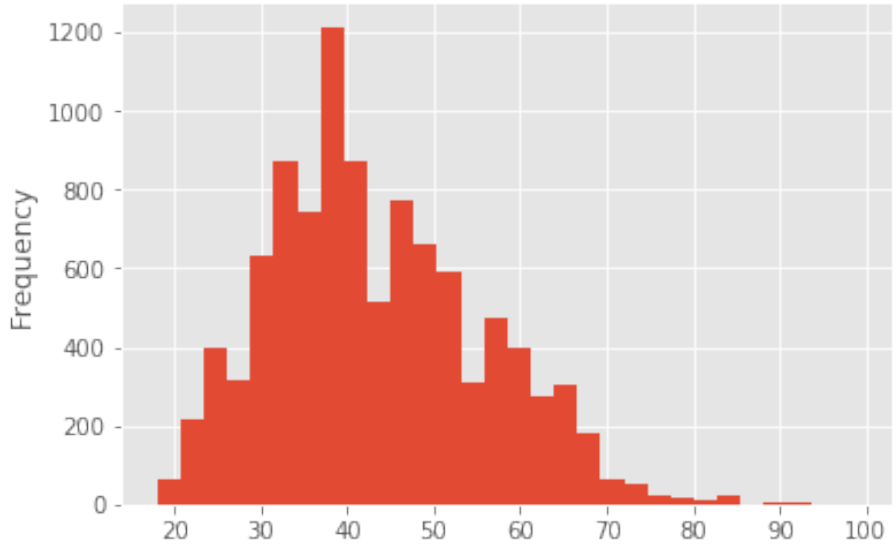

Figura 1. Histograma que mostra a idade das pessoas que fizeram as avaliações

Durante o pré-processamento, as idades das pessoas foram separadas entre categorias (jovens, adultos e terceira idade), o que evidencia adultos como maioria do público femininio que compra roupas online.
Outra característica importante dos dados é que a maioria dos comentários é sobre os produtos e estes bem avaliados, o que mostra que os usuários, na maioria dos casos gostaram bastante dos produtos e a maioria sobre as lojas ficaram neutro. A figura 4 mostra as roupas divididas por categoria e percebe-se que todas estão com boas avaliações médias e com desvio padrão de 1 unidade.
Analisando as figuras 4 e 5, uma das hipóteses (Produto mais bem avaliado é

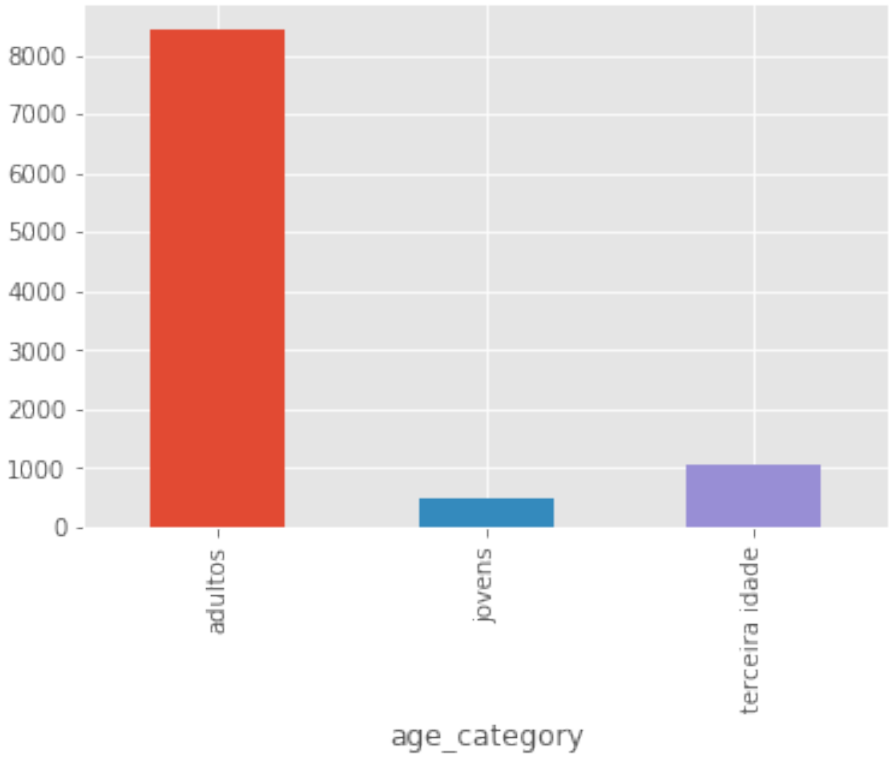

Figura 2. Evidenciando o perfil dos usuários que compram online roupas femininas

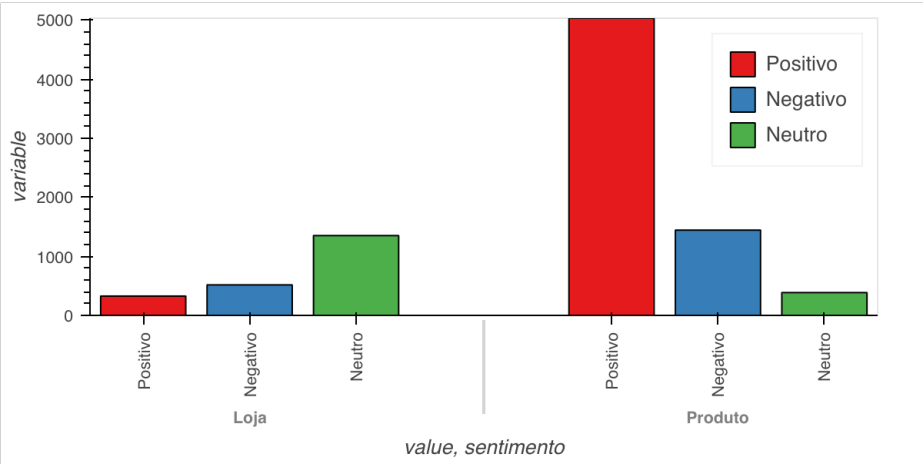

Figura 3. Gráfico com comentários sobre loja e produto separado por sentimento.

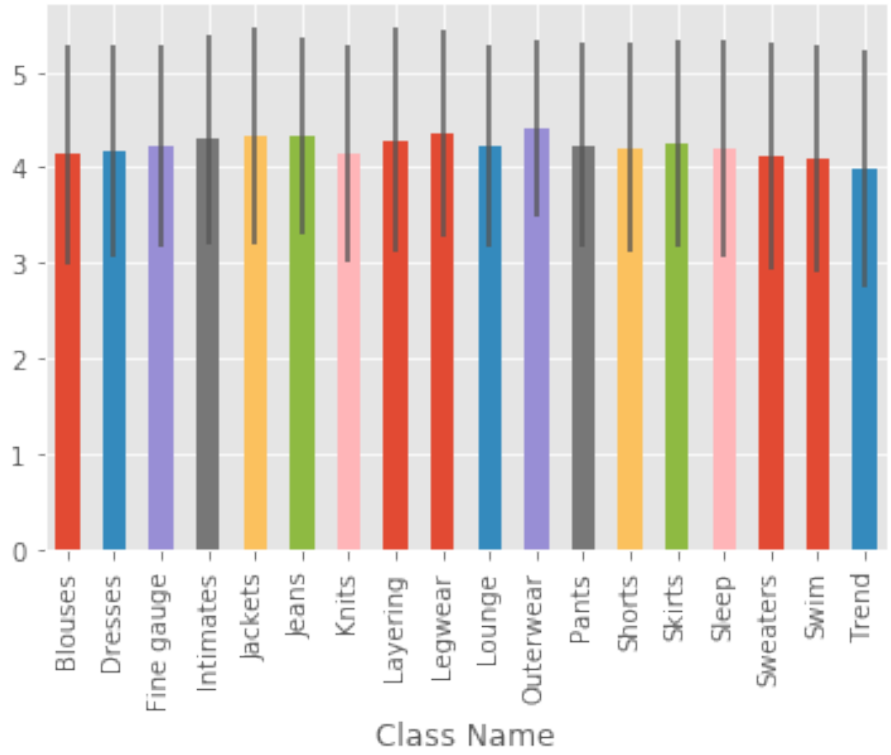

Figura 4. Avaliações das roupas por categoria.

o mais vendido) é inválidada, pois na média todas as categorias de roupas são bem semelhantes.

Figura 5. Quantidade de avaliações por categoria de roupa.

3.2 Mineração de Dados

Para predizer a classe da roupa foi utilizado 3 classificadores, árvore de decisão, KNN e Naive Bayes. Antes de executar o classificador é necessário ajustar as features para categóricas para o classificador conseguir treinar e classificar os dados. Para predizer a classe da roupa todos os classificadores foram bem ruim conseguindo acertar em menos de 30% dos casos:

(i) Naive Bayes: 12,39%
(ii) KNN: 19,97%
(iii) Árvore de decisão: 26,64%

Depois de ver que o classificador era ruim para classficar com o nome da classe do produto, foi testado com outro atributo (Division Name) que tem apenas 4 valores, com ele foi obtido o seguinte resultado:

(i) Naive Bayes: 26,63%
(ii) KNN: 45,94%
(iii) Árvore de decisão: 57,92%

Para tentar melhorar o resultado ainda foi feito uma alteração a feature de rating, alterado de 3 para 5 categorias. Depois desta alteração foi obtido o seguinte resultado:

(i) Naive Bayes: 26,91%
(ii) KNN: 48,50%
(iii) Árvore de decisão: 58,54%

4 Conclusão

Através das técnicas utilizadas nesse projeto, percebeu-se que os adultos compram mais produtos e consequentemente produzem mais feedbacks sobre a compra. Outro ponto a se descatar, é a prefrência dos usuários aos produtos em si em relação a loja, o que leva a projetar que cada vez mais as pessoas se importam menos com marcas. Foi interessante validar uma hipotése (jovens compravam mais vestidos), porém foi visto que a diferença para os outros não era tão grande, outro detalhe é que adultos também preferem comprar vestidos, deixando essa categoria como a mais vendida da base de dados. Não foi possível prever com muita acurácia a classe da roupa comprada, tendo como dados a idade e o review, pois a quantidade de classes de roupas é bastante grande, sendo complicado para o classificador definir qual a categoria pertencia a roupa em questão.