Classificação de Sentimentos em Reviews de Produtos Usando TF-IDF e Naive Bayes

Ellen Beatriz Shen
Insper
São Paulo, Brasil
ellenbs@al.insper.edu.br

João Vitor Magalhães Menezes
Insper
São Paulo, Brasil
joaovmm@al.insper.edu.br

I. Dataset

O dataset utilizado neste trabalho é o "Amazon Reviews Dataset", disponibilizado no Kaggle. Este conjunto de dados contém avaliações textuais de produtos da Amazon, com rótulos binários (`label 1` para avaliações negativas e `label 2` para positivas). A Tabela I resume as características principais do dataset:

TABLE I
Características do Dataset de Avaliações da Amazon

Tipo de Dados	Tamanho	Número de Avaliações
Treino	1,3 GB	3,6 milhões
Teste	170 MB	400 mil

II. Classification Pipeline

A. Pré-processamento de Dados:

- Para a limpeza textual, realizamos a remoção de stopwords, lematização e tokenização. Esse processo garantiu que palavras irrelevantes fossem excluídas, reduzindo o ruído e melhorando a qualidade dos vetores de texto.

B. Vetorização com TF-IDF:

- O modelo de vetorização TF-IDF (Term Frequency-Inverse Document Frequency) foi utilizado para converter os textos das avaliações em vetores numéricos. Esse método atribui um peso a cada palavra com base em sua frequência relativa nos documentos, destacando termos importantes e penalizando aqueles muito comuns.

C. Classificação com Naive Bayes:

- Para a classificação de sentimentos, utilizamos o classificador Naive Bayes Multinomial, adequado para lidar com dados esparsos e textos vetorizados. O treinamento foi realizado em 80% dos dados, enquanto 20% foram utilizados para teste. Esse classificador é conhecido por sua simplicidade e bom desempenho em tarefas de NLP.

III. Evaluation

A Tabela II apresenta as métricas de desempenho do modelo Naive Bayes, calculadas com base nos resultados obtidos sobre o conjunto de teste. As métricas incluem precisão, recall, e F1-Score para cada classe, além da acurácia global e balanceada.

TABLE II
Métricas de Desempenho do Modelo Naive Bayes

Classe	Precisão	Recall	F1-Score
Negativo (1)	0.84	0.85	0.84
Positivo (2)	0.85	0.83	0.84
Acurácia Global	84.0%		
Balanced Accuracy	84.3%		

O modelo atingiu uma acurácia global de 84,0%, com uma acurácia balanceada de 84,3%, mostrando um desempenho equilibrado entre classes. A métrica F1-Score foi de 0,84 para ambas as classes, indicando um bom compromisso entre precisão e recall.

IV. Dataset Size

Para avaliar o impacto do tamanho do dataset no desempenho do modelo, realizamos experimentos variando o número de amostras. Os resultados mostraram que um aumento no número de amostras de treinamento levou a uma melhoria significativa na acurácia até um certo ponto, após o qual os ganhos adicionais foram mínimos. Esse comportamento é esperado, uma vez que, com dados suficientes, o classificador alcança um desempenho estável e não se beneficia de amostras extras.

V. Topic Analysis

Para entender melhor os temas abordados nas avaliações, realizamos uma análise de tópicos utilizando a técnica Latent Dirichlet Allocation (LDA). O objetivo era identificar tópicos latentes e associá-los a diferentes preocupações e elogios dos consumidores.

A. Implementação de LDA

O algoritmo LDA foi aplicado para identificar cinco tópicos principais. Cada avaliação foi representada como uma mistura de tópicos, e cada tópico foi caracterizado como uma distribuição de palavras. O número de tópicos foi definido empiricamente para balancear complexidade e interpretabilidade.

B. Resultados da Análise de Tópicos

A Tabela III apresenta os cinco tópicos identificados e as palavras mais relevantes associadas a cada um deles. Com base nas palavras mais frequentes, os tópicos foram rotulados manualmente.

TABLE III
TÓPICOS IDENTIFICADOS POR LDA E PALAVRAS-CHAVE ASSOCIADAS

Tópico	**Principais Palavras Associadas**
1 - Qualidade do Produto	”produto”, ”qualidade”, ”material”, ”bom”, ”excelente”
2 - Tempo de Entrega	”entrega”, ”rápido”, ”tempo”, ”chegou”, ”atrasado”
3 - Atendimento ao Cliente	”atendimento”, ”suporte”, ”cliente”, ”ajuda”, ”resposta”
4 - Experiência Geral	”experiência”, ”compra”, ”uso”, ”satisfação”, ”valor”
5 - Preço	”preço”, ”custo”, ”barato”, ”caro”, ”valor”

Os resultados indicam que ”Qualidade do Produto” foi o tópico mais frequente, aparecendo em 38% dos reviews, seguido por ”Tempo de Entrega” e ”Atendimento ao Cliente”. Isso sugere que esses são os principais fatores que influenciam a satisfação dos consumidores.

REFERÊNCIAS

1) M. Potthast, ”Sentiment Analysis in Amazon Reviews: A Case Study”, *Proceedings of the 20th IEEE International Conference on Web Information Systems Engineering*, 2019.
2) S. Bird, E. Loper, and E. Klein, ”Natural Language Processing with Python”, O’Reilly Media Inc., 2009.
3) C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
4) ”Amazon Reviews Dataset”, Kaggle: [Amazon Reviews Sentiment Analysis], acessado em Outubro de 2024.
5) L. Guener, ”Sentiment Analysis for Amazon.com Reviews”, ResearchGate, acessado em Outubro de 2024.