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. Introdução

A crescente quantidade de dados textuais gerados na web, como avaliações de produtos e comentários, torna a análise automatizada uma tarefa essencial para entender as preferências dos consumidores. Este projeto visa construir um classificador de sentimentos para reviews de produtos, divididos entre positivos e negativos. Utilizamos um dataset com milhares de avaliações, onde cada texto foi analisado com técnicas de pré-processamento e, posteriormente, classificado por um modelo supervisionado.

II. Dataset

O dataset utilizado neste trabalho foi obtido do Kaggle. Este conjunto de dados consiste em reviews de produtos da Amazon e foi originalmente usado para treinar o modelo de classificação de textos FastText da biblioteca fastText, desenvolvida pelo Facebook. O dataset contém avaliações em inglês classificadas em dois rótulos principais:

- `label 1`: Reviews com classificação negativa.
- `label 2`: Reviews com classificação positiva.

Cada linha do arquivo é composta por um rótulo seguido de um texto representando o review, sem qualquer estrutura adicional.

Os dados são organizados em dois arquivos principais: `train.ft.txt.bz2` e `test.ft.txt.bz2`.

A versão de treino contém 3,6 milhões de avaliações, enquanto o conjunto de teste contém 400 mil avaliações. Neste projeto, utilizamos uma amostra menor do conjunto de teste para simplificar a execução e o desenvolvimento do modelo.

III. Metodologia

O sistema desenvolvido segue uma abordagem clássica de processamento de texto e aprendizado de máquina:

A. Coleta e pré-processamento de dados:

- Os dados foram carregados a partir de um arquivo de texto contendo avaliações de produtos e rótulos. Cada linha foi dividida em rótulo (`label 1` para negativo e `label 2` para positivo) e o conteúdo do review.
- Um DataFrame foi criado utilizando a biblioteca pandas para facilitar a manipulação e visualização dos dados.

B. Tokenização e lematização:

- Aplicamos técnicas de tokenização para dividir cada review em palavras individuais.
- As palavras irrelevantes (stopwords) foram removidas e os termos restantes foram lematizados para reduzir variações morfológicas.

C. Representação Vetorial:

- O texto foi vetorizado usando a técnica TF-IDF (Term Frequency-Inverse Document Frequency), que atribui pesos às palavras de acordo com sua relevância no documento.

D. Classificação:

- Utilizamos o classificador Naive Bayes Multinomial, adequado para textos vetorizados, para categorizar as avaliações como positivas ou negativas.
- O modelo foi treinado em 80 % dos dados e testado nos 20% restantes.

E. Avaliação de Desempenho:

- Os resultados foram avaliados com métricas de precisão, recall e F1-Score, além da acurácia balanceada.

IV. Resultados

A implementação atingiu um desempenho satisfatório com uma acurácia balanceada de 84,3% no conjunto de teste. As principais métricas estão descritas abaixo:

TABLE I
Métricas de Desempenho do Modelo Naive Bayes

Classe Rótulo	**Métricas de Avaliação**		
	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%		

REFERENCES

[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..