Treinar classificador: SVM + bow, SVM + embeddings, BERT + Bônus

Diego Felipe Lourenço da Silva

Especialização Deep Learning

Centro de Informática UFPE

Universidade Federal de Pernambuco

Objetivo Geral:

Analisar sentimentos em avaliações de produtos da Amazon utilizando diferentes modelos de aprendizado de máquina.

Problema: avalie e compare a eficácia de três abordagens

1. SVM com Bag-of-Words.
2. SVM com Embeddings Pré-treinados.
3. BERT para classificação de textos.

Conjunto de dados:

- O conjunto de dados utilizado contém **568.454** avaliações de produtos da Amazon, com informações como texto da avaliação (Text) e nota atribuída (Score).

- Essas notas foram convertidas em três categorias de sentimento: positivo (notas ≥ 4), neutro (nota = 3) e negativo (notas ≤ 2).

- O conjunto de dados é representativo de um amplo conjunto de produtos e foi pré-processado para remover ruídos como números, pontuações e palavras irrelevantes (stopwords).

Metodologia:

1. Pré-processamento:

- Remoção de números, pontuação e stopwords.
- Tokenização.

2. Divisão do Dataset:

- 70% para treinamento.
- 15% para validação.
- 15% para teste.

Modelos Treinados:

1. SVM com Bag-of-Words:

- Textos convertidos em vetores de contagem.

2. SVM com Embeddings:

- Embeddings calculados usando word2vec.

3. BERT:

- Modelo pré-treinado bert-base-uncased.

Desempenho dos Modelos:

Modelo	Acurácia	F1-Pontuação Positiva	F1-Pontuação Neutro	F1-Pontuação Negativa
SVM + Bag-of-Words	77%	87%	18%	41%
SVM + Embeddings	78%	88%	0%	0%
BERT	82%	90%	0%	42%

Conclusão:

1. SVM com Bag-of-Words:

- Simples, rápido, mas menos preciso.

2. SVM com Embeddings:

- Bom equilíbrio entre desempenho e custo computacional.

3. BERT:

- Melhor desempenho geral, mas exige maior poder computacional.

Sugestões:

1. SVM com Bag-of-Words:

- Substituir Bag-of-Words por TF-IDF (Term Frequency-Inverse Document Frequency) para dar mais peso a palavras menos frequentes, mas mais relevantes.

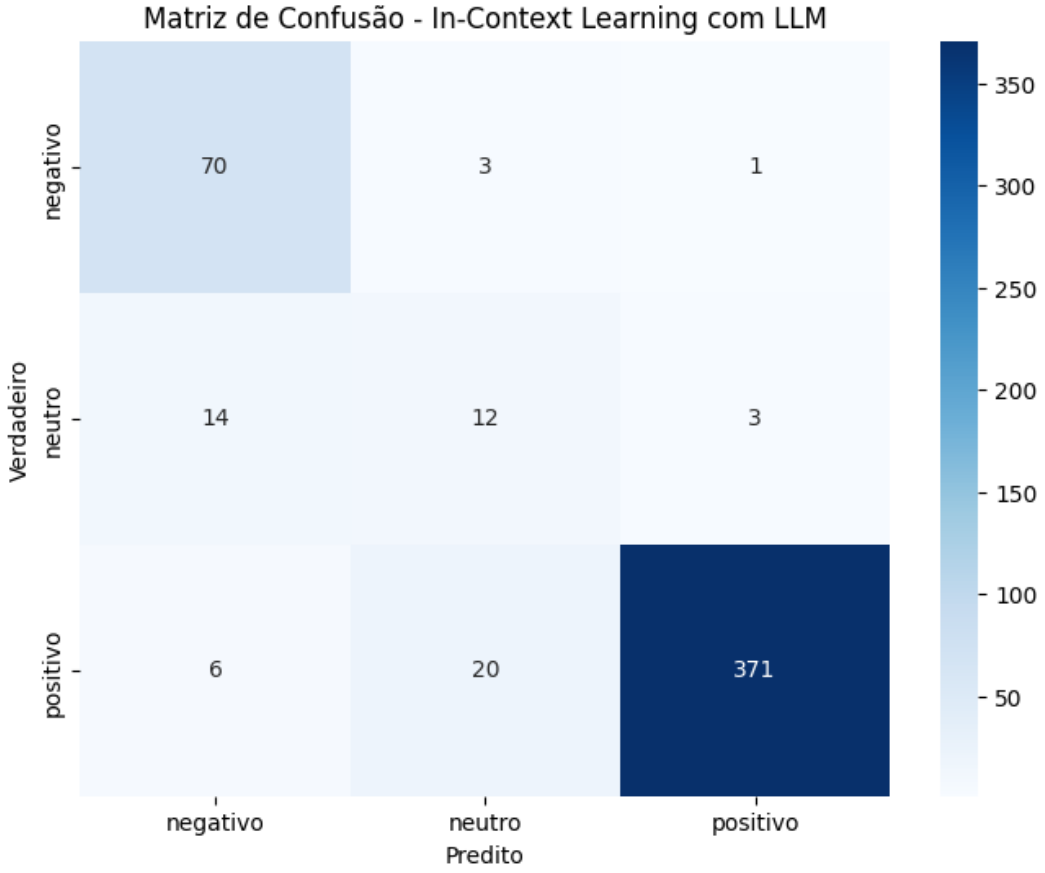
2. SVM com Embeddings:

- Substituir embeddings estáticos como Word2Vec por embeddings contextuais, como aqueles gerados por modelos como BERT ou FastText.

3. BERT:

- Substituir BERT por DistilBERT, uma versão reduzida e mais eficiente:

Bônus:


Matriz de Confusão - In-Context Learning com LLM
Verdadeiro
negativo
neutro
positivo
70
3
1
14
12
3
6
20
371
negativo
neutro
positivo
Predito
350
300
250
200
150
100
50

Referência

[1] A. Rumi, "Amazon Product Reviews Dataset" , Kaggle, [Online]. Disponível em: https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews/data. [Acessado em : 28 de novembro de 2024]