Processamento de Linguagem Natural - PLN

Stefane Adna dos Santos

UFC

PLN - Relembrando

- Utilizada para o desenvolvimento de sistemas computacionais capazes de entender, interpretar e gerar linguagem natural humana.

- O pré-processamento prepara o texto para a vetorização.

UFC

Dataset

- B2W-Reviews01 é um corpus aberto de análises de produtos. Contém mais de 130 mil avaliações de clientes de comércio eletrônico, coletadas no site Americanas.com entre janeiro e maio de 2018.

- O B2W-Reviews01 oferece informações ricas sobre o perfil do avaliador, como gênero, idade e localização geográfica.

UFC

Dataset

- O corpus também tem duas taxas de revisão diferentes:
  - A taxa de escala usual de 5 pontos, representada por estrelas na maioria dos sites de comércio eletrônico.

  - Uma etiqueta "recomende a um amigo": uma pergunta "sim ou não" que representa a vontade do cliente de recomendar o produto a outra pessoa.

UFC

Qual o objetivo do projeto?

UFC

Análise

1. O gênero dos usuários influência no sentimento das avaliações?
2. A idade dos usuários influência na quantidade e no sentimento das avaliações?
3. Quais são os produtos com as melhores e piores avaliações?
4. Quais categorias de produtos são mais avaliadas?

UFC

Limpeza e exploração

- Remover dados duplicados.
- Remover dados nulos.
- Altera o tipo de algumas colunas.
- Transforma os dados de categoria para o tipo numérico.
- Qual coluna utilizar como rótulo dos textos.

UFC

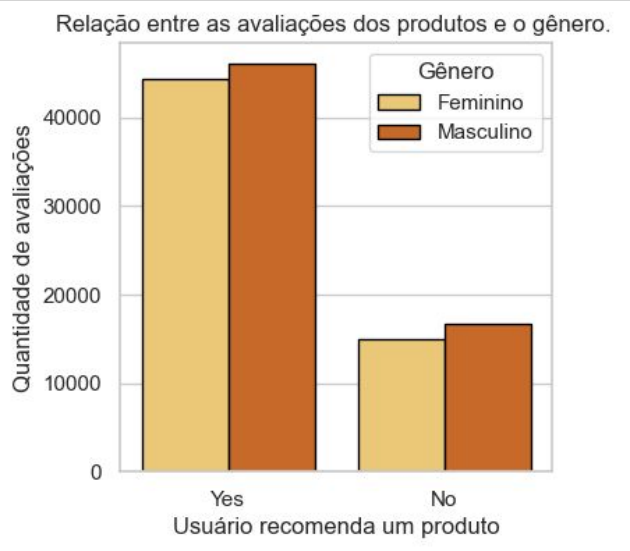
Limpeza e exploração

- O que são rótulos?
  - Os rótulos (ou "labels" em inglês) são as informações que se deseja prever ou classificar a partir dos dados de treinamento. Em outras palavras, são as respostas ou resultados conhecidos usados para ensinar um modelo de Machine Learning a fazer previsões precisas.

UFC

Gráficos

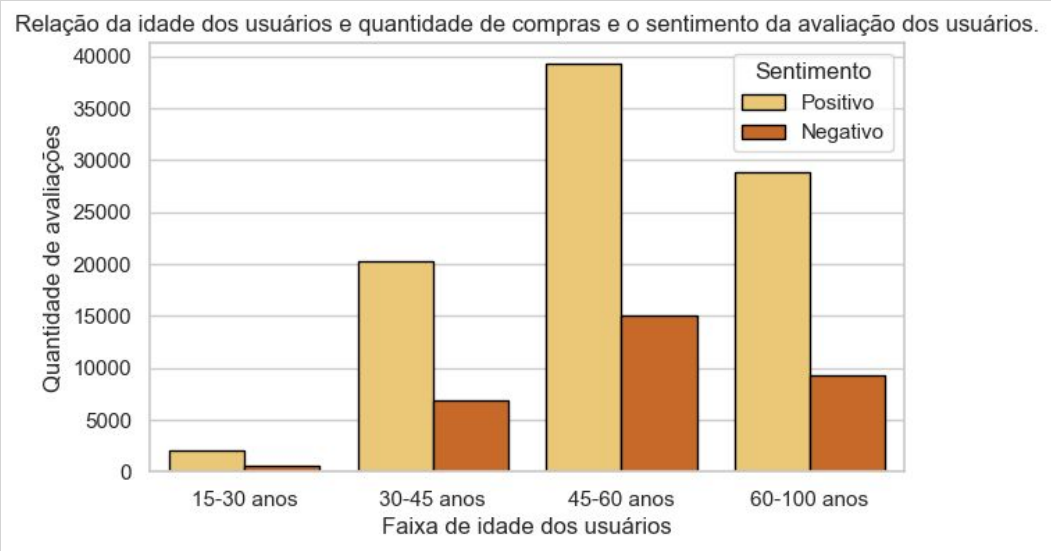
1. O gênero dos usuários influência no sentimento das avaliações?



UFC

Gráficos

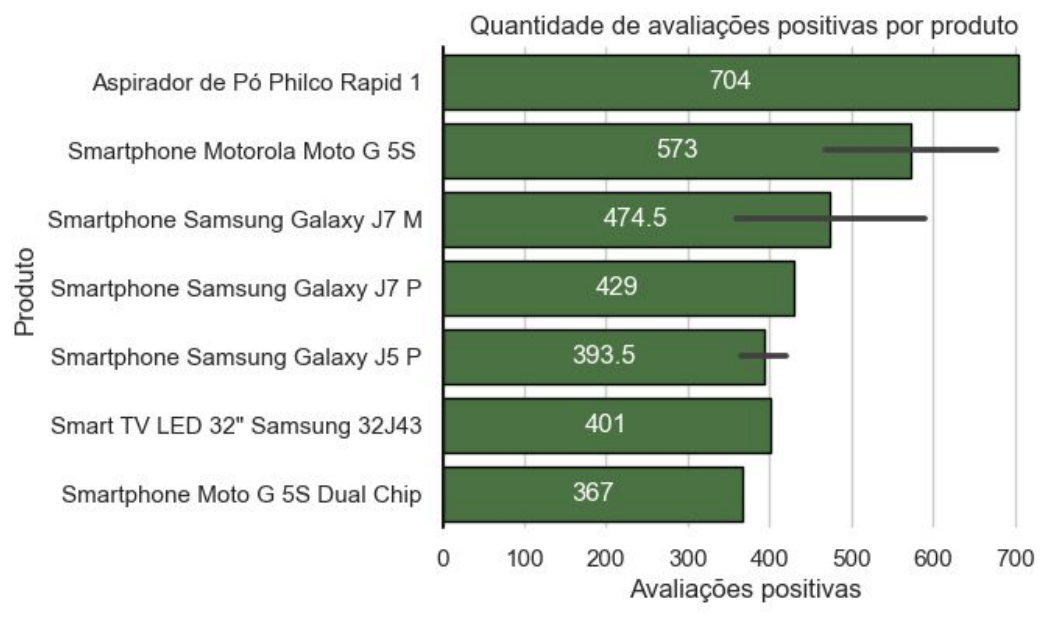
2. A idade dos usuários influência na quantidade e no sentimento das avaliações?



UFC

Gráficos

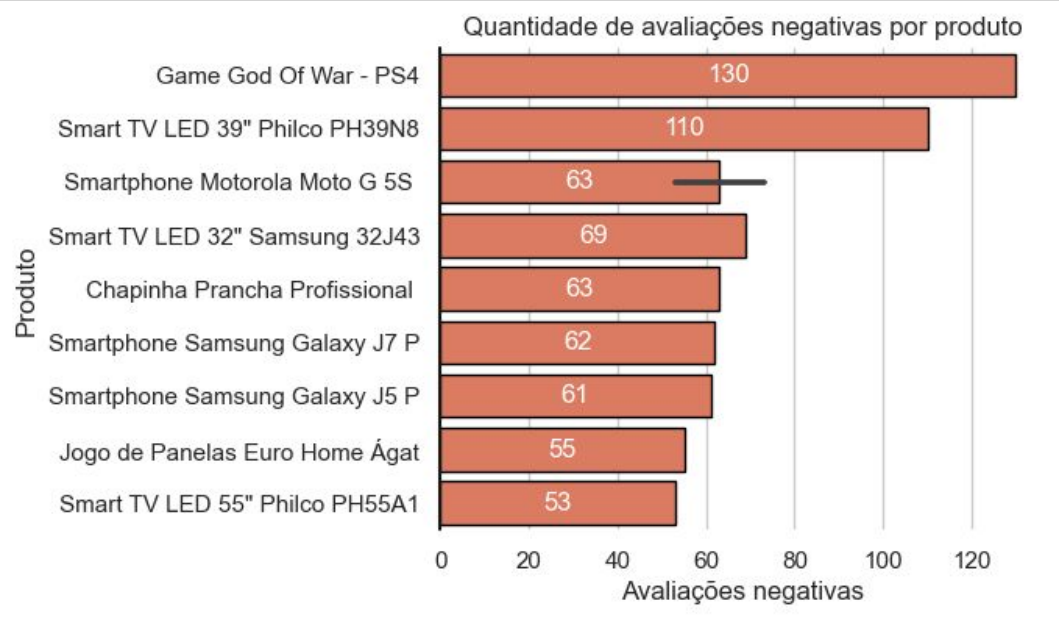
3. Quais são os produtos com as melhores e piores avaliações?



UFC

Gráficos

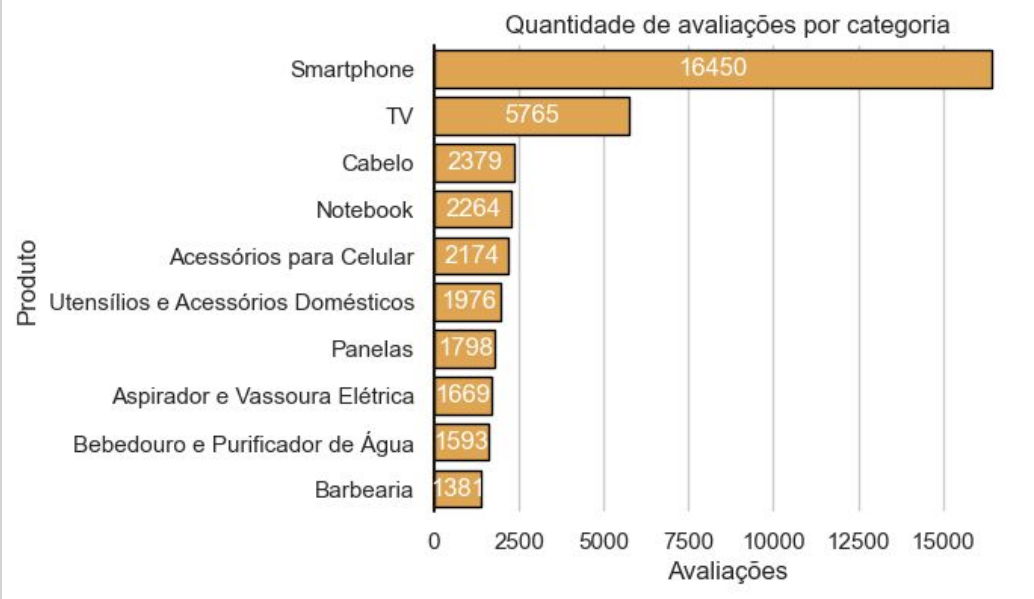
3. Quais são os produtos com as melhores e piores avaliações?



UFC

Gráficos

4. Quais categorias são mais avaliadas?



UFC

Gráficos

4. Nuvem de Palavras

UFC

Pré-processamento textual

UFC

Pré-processamento

- Remoção de caracteres especiais, números e pontuações.
- Tokenização e remoção de StopWords.
- Stematização e lematização.
- Necessidade de escolha do melhor pipeline de pré-processamento.

UFC

Experimentos

UFC

Qualidade de software

- Evitar códigos duplicados.
- Modularização.
- Estruturação do código.
- Continuidade.
- Documentação.
- Inteligibilidade.

UFC

Treino e teste

- A principal importância dessa divisão é avaliar o desempenho do modelo em dados que ele nunca viu antes, o que é essencial para avaliar a capacidade do modelo de generalizar para novos dados.

UFC

Métricas de validação

- As métricas de validação são usadas em Machine Learning para avaliar a capacidade de um modelo de fazer previsões precisas e confiáveis.

Texto	Rótulo Real	Rótulo Previsto
Eu amei este produto.	Avaliação positiva	Avaliação positiva
Eu odiei este produto.	Avaliação negativa	Avaliação negativa
Este produto é muito ruim.	Avaliação negativa	Avaliação positiva

UFC

Métricas de validação

- **Verdadeiro Positivo (True Positive - TP):** é quando o modelo classifica corretamente uma amostra positiva como positiva.

- **Verdadeiro Negativo (True Negative - TN):** é quando o modelo classifica corretamente uma amostra negativa como negativa.

UFC

Métricas de validação

- **Falso Negativo (False Negative - FN):** é quando o modelo classifica incorretamente uma amostra positiva como negativa.

- **Falso Positivo (False Positive - FP):** é quando o modelo classifica incorretamente uma amostra negativa como positiva.

UFC

Métricas de validação

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Figura: Matriz de confusão

UFC

Métricas de validação

- **Acurácia (Accuracy):** É a proporção de previsões corretas feitas pelo modelo em relação ao número total de previsões.

- **Precisão (Precision):** É a proporção de previsões positivas corretas feitas pelo modelo em relação ao número total de previsões positivas.

UFC

Métricas de validação

- **Recall ou Sensibilidade (Recall/Sensitivity):** É a proporção de previsões positivas corretas feitas pelo modelo em relação ao número total de valores reais positivos.

- **F1 Score:** É a média harmônica entre a precisão e o recall. É uma medida mais balanceada que a acurácia, quando os dados são desbalanceados.

UFC

Vetorização

- Qual a melhor vetorização?
  - TF-IDF.
  - Doc2Vec.

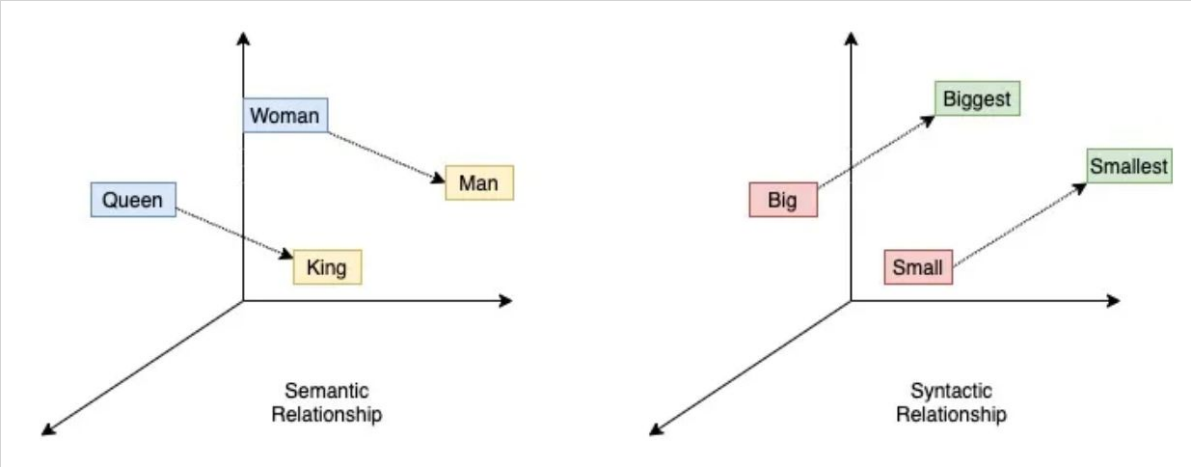


Figura: Word2Vec

UFC

Vetorização

- Qual a melhor vetorização?
  - Acurácia utilizando LogisticRegression e TFIDF: **90.19%**
  - Acurácia utilizando LogisticRegression e Doc2Vec: **85.8%**

UFC

Vetorização

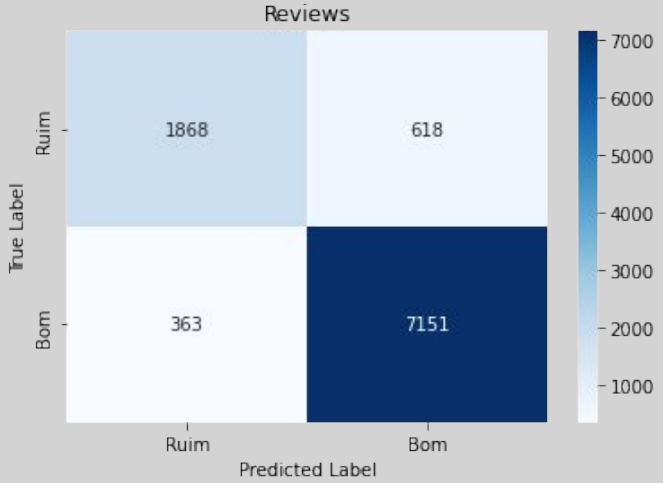


Figura: TF-IDF

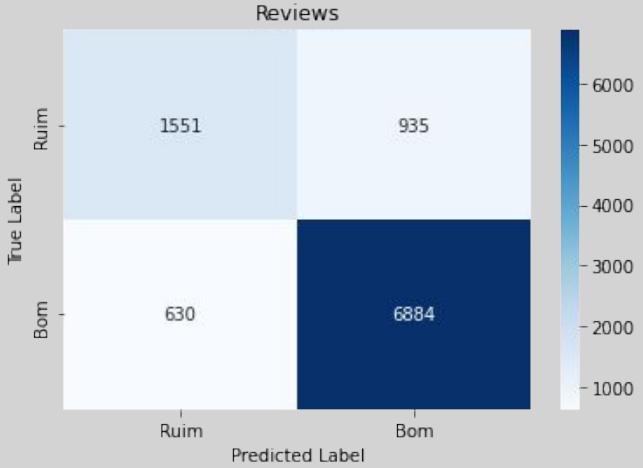


Figura: Doc2Vec

UFC

Classificadores

- Qual o melhor classificador?
  - Floresta Aleatória.
  - Bernoulli
  - KNN.
  - Regressão Logística.
  - Bagging Classifier.

- Existem inúmeros classificadores:
  - https://scikit-learn.org/stable/supervised_learning.html

UFC

GridSearch

- O Grid Search é uma técnica de busca de hiperparâmetros usada em Machine Learning para encontrar a melhor combinação de valores para os parâmetros de um modelo.

UFC

GridSearch

- O Grid Search funciona criando uma grade de valores para cada hiperparâmetro que se deseja ajustar e testando todas as combinações possíveis desses valores para encontrar a combinação que resulta no melhor desempenho do modelo em um conjunto de dados de validação.

UFC

Validação cruzada

- A validação cruzada é uma técnica em Machine Learning usada para avaliar a capacidade de um modelo de generalizar para novos dados.

- É uma técnica de validação de modelos que consiste em dividir os dados em vários conjuntos de treinamento e teste, e realizar testes repetidos para verificar a variabilidade do modelo.

UFC

Validação cruzada

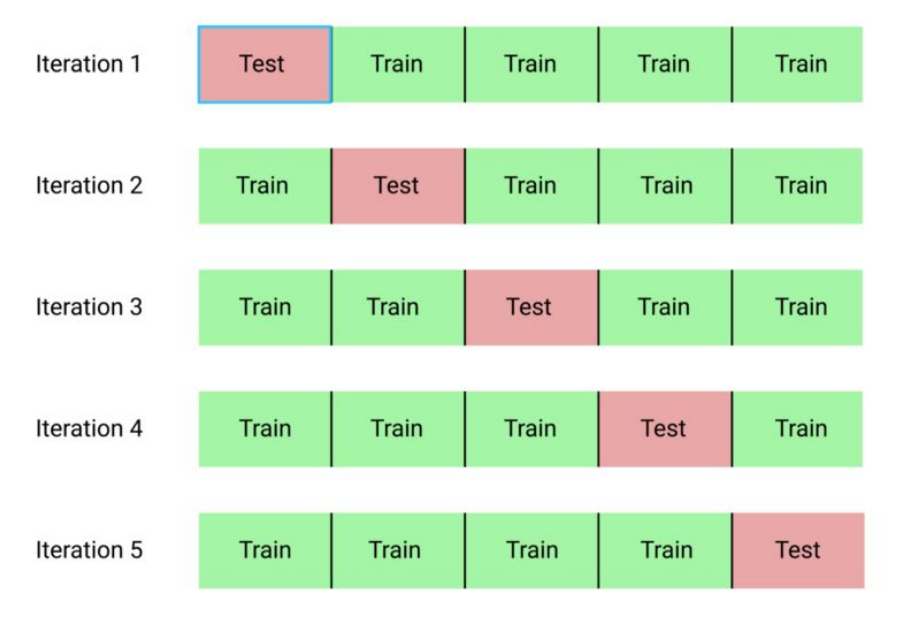


Figura: K-Fold

UFC

Validação cruzada

Classificador	Acurácia	F1Score
Logistic Regression	**90.32 ± 0.43**	**93.65 ± 0.29**
BernoulliNB	86.12 ±0.52	90.90 ± 0.36
KNN	75.02 ± 6.00	83.38 ± 7.45
Random Forest	75.31 ± 0.06	85.89 ± 0.03
Bagging Classifier	90.25 ± 0.36	93.61 ± 0.25

Tabela: Resultado da validação cruzada

UFC

Pré-processamento

Processamento e texto	Acurácia
review_text + stemização	90.19%
review_text + lematização	90.21%
review_text + lematização + stemização	89.93%
review_title + stemização	88.79%
review_title + lematização	89.05%
review_title + lematização + stemização	88.78%
review_text +review_title + stemização	91.96%
review_text +review_title + lematização	**92.49%**
review_text +review_title + lematização + stemização	92.17%

Tabela: Resultado dos testes para escolher o pipeline do pré-processamento

UFC

Pipeline escolhido

- Os melhores resultados foram obtidos utilizando os algoritmos de lematização, TF-IDF e o classificador de Regressão Logística. Este algoritmos em conjunto obtiveram **92.49%** de acurácia nos testes.

UFC

Pipeline

- Foram desenvolvidos dois pipelines, sendo uma para treinamento do modelo e outro para predizer um texto utilizando o modelo.

UFC

Reprodutividade

- A reprodutibilidade se refere à capacidade de executar um conjunto de instruções (código) e obter os mesmos resultados toda vez que o código é executado, independentemente da plataforma ou ambiente em que está sendo executado.

UFC

Reprodutividade

- A reprodutibilidade de código é importante porque permite que outras pessoas verifiquem e validem os resultados, tornando a ciência mais transparente e confiável. Também ajuda a evitar problemas comuns, como erros de arredondamento ou variações na execução devido a diferentes versões de bibliotecas.

UFC

Reprodutividade

- Criar requirements.txt
  - pip freeze > requirements.txt

- Para utilizar este projeto deve-se clonar este repositório e executar o seguinte comando dentro da pasta do projeto:
  - pip install -r requirements.txt

UFC

Reprodutividade

- Para treinar um novo modelo pode-se executar o comando abaixo:
  - python src/americanas/train.py

- Para realizar a predição de um texto, pode-se executar o comando abaixo:
  - python src/americanas/predict.py --text "este produto é muito bom"

UFC

Streamlit

- Streamlit é um framework de código aberto para criar aplicativos da web de dados em Python. Ele permite aos usuários criar aplicativos interativos em questão de minutos com um código Python simples e fácil de entender.

UFC

Streamlit

- Com o Streamlit, é possível criar aplicativos interativos para visualização e análise de dados, prototipagem de modelos de aprendizado de máquina, dashboard de monitoramento e muito mais.

UFC

Streamlit

Figura: Streamlit

UFC

Referencias

- Github. **Análise de Sentimento Americanas**. Disponível em: < https://github.com/stefaneadna/nlp_sentiment_analysis_neoway>. Acesso em 04 de abril de 2023.

UFC