



Camilly Alves RM 550210

João Vitor Martins RM 98744

Murilo Krauss RM 98262

Rafael Lima RM 88444

Challenge Sprint 1 – 2024

Processamento de Linguagem Natural, Chatbots & Virtual Agents

Prof Leonardo Ruiz Orabona

Para desenvolver um algoritmo de análise de sentimentos, utilizamos dois datasets: um para treinamento e outro para validação. O dataset de teste, contendo 100 linhas, possui as seguintes colunas:


- Serviço
- Comentário
- Sentimento

As colunas "Serviço" e "Sentimento" foram incluídas para que o algoritmo pudesse aprender a interpretar tanto o sentimento quanto o serviço, independentemente do comentário. No dataset de validação, composto por 104 registros, incluímos as colunas:

- Serviço
- Comentário

Primeiramente, realizamos uma análise descritiva simples das bases para entender o comportamento delas, a distribuição das informações e o tipo de dados contidos nelas:

#### Dataset de treino

✓ 0s  `df.sample(5)`

	servico	comentario	sentimento
22	Emissão de carteira de trabalho	Precisei de suporte técnico para continuar.	negativo
57	Emissão de carteira de trabalho	Todo o procedimento foi muito claro.	positivo
94	Abertura de empresa	Houve muitos erros no site.	negativo
59	Abertura de empresa	Precisei de ajuda e fui prontamente atendido.	positivo
98	Licenciamento de veículo	Foi fácil e rápido licenciar meu veículo.	positivo

✓ 0s [5] `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   servico      100 non-null    object
1   comentario   100 non-null    object
2   sentimento   100 non-null    object
dtypes: object(3)
memory usage: 2.5+ KB
```

✓ 0s [6] `df.describe()`

	servico	comentario	sentimento
count	100	100	100
unique	5	97	2
top	Emissão de RG	O site ficou fora do ar várias vezes.	positivo
freq	20	2	53

## ✓ Dataset de validação

✓ [7] `df_teste.sample(5)`

	servico	comentario
3	Emissão de RG	Fui muito bem atendido ao solicitar a emissão...
93	Emissão de RG	O processo de emissão do RG pelo site foi nor...
79	Licenciamento de veículo	A experiência de licenciar meu veículo pelo s...
37	Licenciamento de veículo	O processo de licenciamento pelo site foi mui...
89	Abertura de empresa	O suporte para abrir minha empresa pelo site ...

✓ [8] `df_teste.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 104 entries, 0 to 103
Data columns (total 2 columns):
#   Column      Non-Null Count  Dtype
---  -
0   servico     104 non-null   object
1   comentario  104 non-null   object
dtypes: object(2)
memory usage: 1.8+ KB
```

✓ [9] `df_teste.describe()`

	servico	comentario
count	104	104
unique	3	77
top	Emissão de RG	Tive muitos problemas ao tentar abrir minha e...
freq	35	3

Após essa breve análise, partimos para o desenvolvimento do algoritmo. Inicialmente, os comentários foram convertidos para letras minúsculas, tiveram as stopwords removidas, foram tokenizados e lematizados. Montamos o algoritmo como um pipeline e utilizamos GridSearch para localizar os parâmetros ideais para o modelo.

Depois de ajustado o modelo, visualizamos as informações do dataset de validação. Algumas visualizações simples e rápidas que fizemos incluem:

- Qual foi o serviço mais comentado? R: Emissão de RG
- Qual foi o serviço com mais comentários positivos? R: Abertura de empresas
- Qual foi o serviço com mais comentários negativos? R: Emissão de RG

```
# Análise e visualização
estatisticas = df_teste['sentimento_predito'].value_counts()
comentarios_repetidos = Counter(df_teste['comentario']).most_common()

servico_positivo = df_teste[df_teste['sentimento_predito'] == 'positivo']['servico'].mode()[0]
servico_negativo = df_teste[df_teste['sentimento_predito'] == 'negativo']['servico'].mode()[0]
servico_mais_comentado = df_teste['servico'].mode()[0]

sentimentos_por_servico = df_teste.groupby(['servico', 'sentimento_predito']).size().unstack().fillna(0)
servico_mais_positivo = sentimentos_por_servico['positivo'].idxmax()
servico_mais_negativo = sentimentos_por_servico['negativo'].idxmax()

print(f'Serviço com mais comentários: {servico_mais_comentado}')
print(f'Serviço mais positivo: {servico_mais_positivo}')
print(f'Serviço mais negativo: {servico_mais_negativo}')
```

↗ Serviço com mais comentários: Emissão de RG  
 ↗ Serviço mais positivo: Abertura de empresa  
 ↗ Serviço mais negativo: Emissão de RG

Em seguida, identificamos os comentários mais repetidos, o que nos fornece uma ideia clara do que está satisfatório ou problemático para os clientes.

↗ Tabela de Comentários Mais Relevantes - Positivos:

	Comentário	Quantidade	Sentimento	Acurácia do Modelo
0	Fiquei impressionado com a facilidade de real...	3	Positivo	0.85
1	O licenciamento do meu veículo pelo site foi ...	3	Positivo	0.85
2	Tive muitos problemas ao tentar abrir minha e...	3	Positivo	0.85

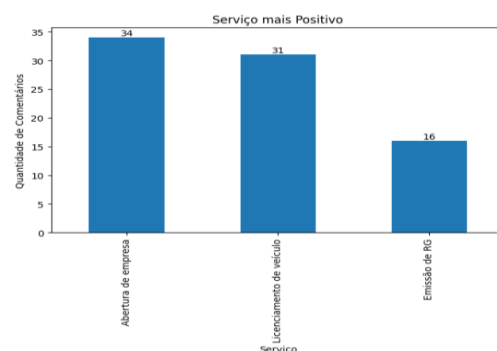
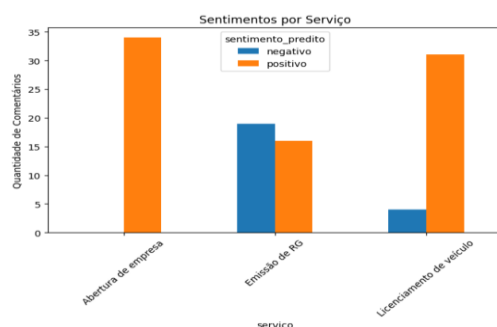
---

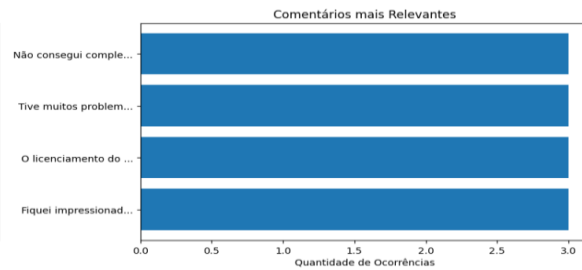
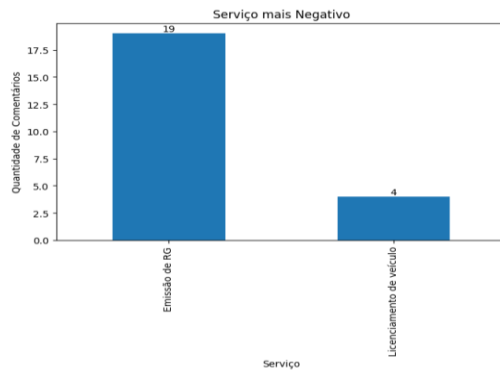
↗ Tabela de Comentários Mais Relevantes - Negativos:

	Comentário	Quantidade	Sentimento	Acurácia do Modelo
0	Não consegui completar o processo de emissão ...	3	Negativo	0.85
1	O atendimento para emissão do RG pelo site fo...	3	Negativo	0.85
2	Não recomendo o licenciamento de veículo pelo...	3	Negativo	0.85

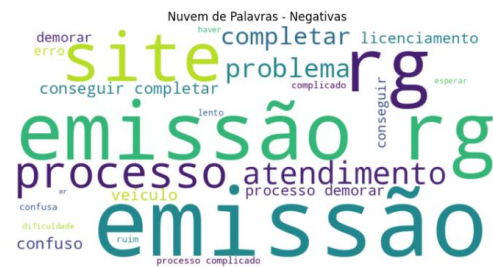
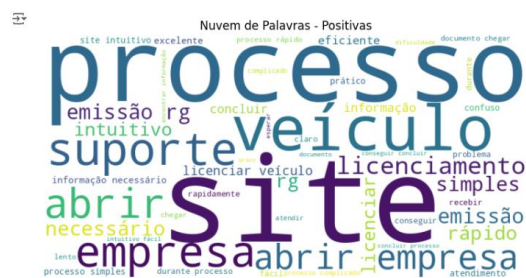
Para uma compreensão mais ampla, montamos quatro gráficos diferentes, que nos ajudam a entender melhor a situação:

- Sentimentos por Serviço
- Serviço mais positivo
- Serviço mais negativo
- Comentários mais frequentes





Por fim, para entender melhor os comentários, criamos nuvens de palavras, ideais para visualizar as principais palavras dos comentários, tanto negativos quanto positivos.



Essas visualizações nos permitem identificar rapidamente as palavras mais frequentemente associadas a sentimentos positivos e negativos nos comentários, proporcionando insights valiosos sobre a "saúde" do serviço oferecido.