Relatório – Recuperação da Informação

Aluno: Pablo Henrique Silva de Faria
Matrícula: 202235012

1. Introdução

A análise de posicionamentos políticos em textos é um problema no contexto atual, devido ao fato de que grande parte da comunicação ocorre em meios digitais. Redes sociais, veículos de imprensa e blogs produzem diariamente um grande volume de informações que podem carregar polarizações ideológicas explícitas ou implícitas, influenciando a leitura e o estudo de diversos grupos de pessoas.

Nesse cenário, ferramentas automáticas de classificação de texto podem auxiliar pesquisadores em ciência política, jornalistas e desenvolvedores de sistemas de monitoramento a identificar tendências de opinião e padrões discursivos, além de colaborar para a leitura menos parcial para pessoas menos desconfiadas, ou seja, ferramentas desse tipo podem ser usadas como alertas ao ler textos tendenciosos. O problema abordado neste trabalho consiste na **classificação da polaridade política de textos curtos em português** em duas classes: *-1* (esquerda política) e *+1* (direita política), mostrando o quanto cada texto se aproxima dos extremos, ou seja, a **classificação demonstrará um número entre -1 e 1, mas não necessariamente igual a -1 ou 1.** Em outras palavras, o quanto um texto pode ser dito “de direita” ou “de esquerda”.

O objetivo principal é **avaliar o desempenho de diferentes modelos de aprendizado de máquina** — neste caso, **Naive Bayes** e **Regressão Logística** — aplicados ao problema. Além da avaliação quantitativa, buscou-se compreender a interpretabilidade dos modelos por meio das técnicas **SHAP** e **LIME**, que permitem identificar as palavras mais relevantes para cada predição.

2. Materiais e Método

2.1 Base de dados

A base de dados utilizada foi construída utilizando-se web-scrapers feitos em Python com o uso de bibliotecas como **cloudscraper** e **bs4** criados para obter notícias de determinados sites com posicionamentos políticos conhecidos, sendo eles O Antagonista (direita), CartaCapital (esquerda), Jovem Pan (direita) e Diário do Centro do Mundo (esquerda). Os textos foram rotulados de forma binária, atribuindo-se a cada instância um valor **-1** ou **+1**, de acordo com a posição política expressa.

No total, a base contém **399 textos**, distribuídos de forma relativamente balanceada entre as duas classes:

- Classe -1 (esquerda): 200 notícias
- Classe +1 (direita): 199 notícias

Essa distribuição equilibrada é importante, pois reduz o risco de enviesamento do modelo em relação a uma classe específica. A base foi dividida em **75% para treino, 15% para validação e 15% para teste**.

2.2 Experimento e configuração

O processo de pré-processamento incluiu:

- **Tokenização** dos textos;
- **Remoção de stopwords** em português (NLTK);
- Representação vetorial utilizando **Bag-of-Words** (CountVectorizer) e **TF-IDF**.

Dois algoritmos foram avaliados:

1. **Naive Bayes Multinomial (MultinomialNB)** — com parâmetros padrão da biblioteca Scikit-learn.
2. **Regressão Logística (Logistic Regression)** — com solver *liblinear* e parâmetro de regularização C=1.0.

A avaliação dos modelos foi realizada utilizando as métricas:

- **Acurácia**
- **Precisão**
- **Revocação**
- **F1-score**

Por fim, aplicaram-se as técnicas **SHAP** e **LIME** para interpretar os resultados, identificando as palavras que mais influenciaram as decisões de cada modelo.

3. Resultados e Discussão

A Tabela 1 apresenta os resultados obtidos pelos dois classificadores no conjunto de teste.

Tabela 1 – Desempenho dos modelos na base de teste para cada classe

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
Naive Bayes (**direita**)	0.78	0.75	0.75	0.75
Naive Bayes (**esquerda**)	0.78	0.80	0.80	0.80
Regressão Logística (**direita**)	0.89	1.00	0.75	0.86
Regressão Logística (**esquerda**)	0.89	0.83	1.00	0.91

A análise mostra que ambos os modelos obtiveram desempenho satisfatório, com a Regressão Logística apresentando resultados ligeiramente superiores em **todas as métricas**. Isso pode ser explicado pelo fato de a Regressão Logística conseguir capturar melhor relações lineares entre os atributos derivados da representação vetorial **TF-IDF**.

Também é possível observar algumas diferenças devidos às classes: para a classe **“direita”**, a Regressão Logística teve **Precisão = 1.00**, mas **Recall = 0.75**, demonstrando que o modelo foi muito “cauteloso”, ou seja, quando ele classifica algo como “direita”, quase nunca erra, mas deixa escapar algumas amostras dessa classe. Para a classe **“esquerda”**, aconteceu o oposto: **Recall = 1.00** (nenhuma amostra de esquerda foi perdida), mas **Precisão = 0.83** (houve mais falsos positivos).

Essa diferença pode estar relacionada à distribuição dos termos no corpus: talvez notícias de “esquerda” compartilhem vocabulário mais difuso, enquanto as de “direita” têm termos mais específicos e discriminativos.

Do ponto de vista interpretativo, os resultados de SHAP e LIME indicaram que determinadas palavras-chave, associadas a ideologias ou partidos específicos, possuem grande peso nas classificações. Por exemplo, termos como **“Quaest”, “Aprovação”** e **“Aponta”** foram mais associados a notícias de **direita**, dado que o **LIME** os atribuiu valores positivos, ao passo que palavras como **“Lula”, “Governo”, “Pesquisa”** e **“Nova”** foram associadas a notícias de **esquerda,** possivelmente indicando que pesquisas e “novidades” podem estar mais associadas ao **governo atual**. O SHAP mostrou também que as palavras mais relevantes individualmente para a classificaçao das notícias foram **“Quaest”, “Nova”, “Pesquisa”, “Aprovação”, “Lula”,** e **“Aponta”**. Quanto às palavras **“Bolsonaro”**, e **“Lula”**, **“Bolsonaro”** foi uma palavra forte, mas não tanto quanto **“Lula”**, sendo 0.07 ponto inferior na classificaçao do **SHAP**, o que sugere que notícias sobre Lula foram mais frequentes ou mais padronizadas no *dataset*, tornando sua presença um forte indício de esquerda. Bolsonaro, embora relevante, não teve o mesmo peso. Isso pode ser discutido como possível **viés de cobertura jornalística**.. Veja a figura 1.

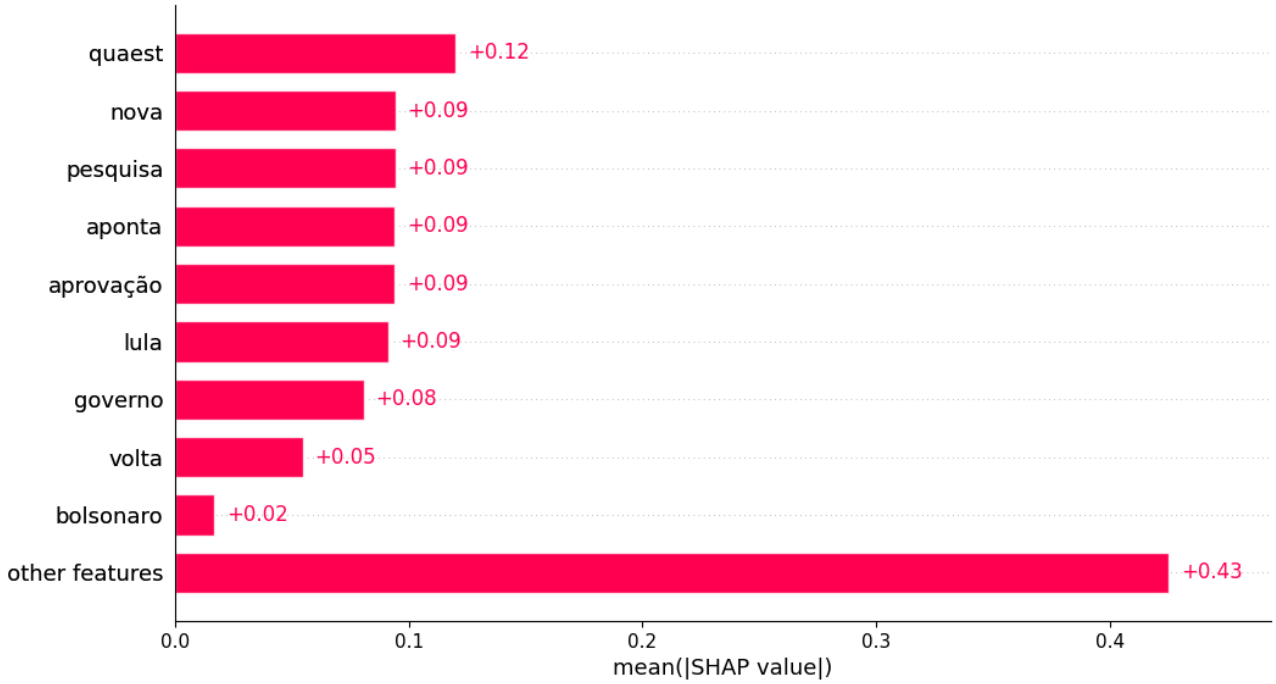

Figura 1.

4. Propostas de melhoria

Embora a Regressão Logística tenha superado o Naive Bayes em todas as métricas, a análise crítica revela que o ganho de desempenho está fortemente associado a termos específicos do corpus. O

peso de palavras como *‘Lula’* e *‘Quaest’* na identificação de notícias de esquerda sugere um viés lexical que pode comprometer a generalização do modelo. Além disso, o *trade-off* entre precisão e *recall* nas duas classes mostra que, embora o modelo seja consideravelmente confiável ao identificar notícias de direita, ele falha em recuperar todas as instâncias dessa classe. Isso indica que, apesar de satisfatórios, os resultados devem ser **interpretados com cautela**, pois o desempenho pode não se manter em contextos menos enviesados ou em dados de maior diversidade lexical. Apesar dos resultados satisfatórios, este trabalho apresenta limitações que podem ser superadas em estudos futuros:

- **Ampliação da base de dados**: o tamanho reduzido (399 exemplos) limita a generalização. A coleta de milhares de textos tornaria os modelos mais robustos.
- **Uso de embeddings semânticos**: representações densas como **Word2Vec**, **FastText** ou **BERTimbau** poderiam capturar relações semânticas mais complexas.
- **Exploração de outros algoritmos**: como Máquinas de Vetores de Suporte (SVM), Random Forest ou redes neurais profundas.
- **Refinamento das explicações**: aplicação mais extensa de SHAP/LIME para diversos exemplos, permitindo uma análise qualitativa mais rica.

5. Conclusão

Esse trabalho implementou e avaliou dois algoritmos clássicos de aprendizado de máquina — Naive Bayes e Regressão Logística — aplicados à classificação de polaridade política em notícias em português. Os resultados demonstraram que ambos os modelos foram eficazes, com ligeira vantagem para a Regressão Logística. A aplicação de técnicas de interpretabilidade (SHAP e LIME) permitiu compreender como os modelos tomaram suas decisões, revelando as palavras mais relevantes em cada predição.

De forma geral, a pesquisa evidenciou o potencial de métodos de PLN aplicados ao contexto político, bem como as limitações de trabalhar com bases pequenas e representações simples. As propostas de melhoria indicam caminhos para trabalhos futuros que possam alcançar resultados ainda mais expressivos e confiáveis.