



TECHNICAL REPORT

Aluno: Pedro Vinicius Félix Rosa Viana

1. Introdução

O conjunto de dados utilizado neste trabalho foi construído a partir de um sistema de coleta automatizada que extrai, a cada 30 minutos, manchetes de diversos portais de notícias brasileiros, com base em um conjunto de palavras-chave relacionadas a políticos influentes no cenário nacional. As manchetes originais, em português, são traduzidas para o inglês e submetidas a análises de sentimento — tanto na versão original quanto traduzida — utilizando a API de Linguagem Natural do Google.

Atualmente, o dataset contempla manchetes provenientes de veículos de diferentes perfis editoriais, como UOL, Folha de S. Paulo, G1, R7, O Antagonista, Senso Incomum e Terça Livre. As palavras-chave incluem nomes de presidentes, ex-presidentes, governadores, ministros, ex-ministros e potenciais candidatos aos cargos de prefeito em São Paulo e no Rio de Janeiro.

Neste trabalho, utilizamos esse dataset para explorar tarefas de Processamento de Linguagem Natural (PLN), incluindo pré-processamento textual, extração de atributos e análise comparativa de modelos de classificação. O objetivo é avaliar como diferentes abordagens de representação textual podem influenciar o desempenho da análise de sentimento.

O dataset original contém diversas variáveis, mas duas possuem maior relevância para os objetivos deste estudo: *headlinePortuguese*, que representa o texto da manchete em português, e *sentimentScorePortuguese*, que corresponde ao escore de sentimento atribuído pela API de Linguagem Natural do Google.

Antes da extração dos atributos, o texto em *headlinePortuguese* passou por um conjunto de procedimentos de pré-processamento, necessários para padronizar a estrutura linguística e reduzir ruídos. As etapas aplicadas foram:

- **Normalização:** conversão de todos os caracteres para minúsculas, remoção de acentos e padronização de espaços em branco.
- **Remoção de URLs e menções:** manchetes que continham endereços de sites ou menções diretas foram filtradas para evitar interferências na extração de



características textuais.

- **Stopwords:** aplicação de listas de palavras irrelevantes em português, removendo termos funcionais que não contribuem para a análise de sentimento.
- **Stemming:** redução das palavras às suas raízes utilizando algoritmos clássicos, com o objetivo de diminuir a variação morfológica.
- **Lemmatização:** transformação das palavras em suas formas canônicas, permitindo manter relações semânticas mais próximas do texto original.

Após o pré-processamento, foram aplicadas diferentes técnicas de extração de atributos, representando combinações variadas de granularidade e abstração semântica. Os extratores utilizados foram:

2. **Estatísticas individuais:** medidas simples, como contagem de tokens, média de tamanho das palavras e frequência de caracteres específicos.
3. **CountVectorizer:** representação baseada em bag-of-words, transformando o texto em vetores de frequência de termos.
4. **TF-IDF (Term Frequency–Inverse Document Frequency):** ponderação dos termos baseada em relevância relativa, permitindo reduzir o peso de palavras muito frequentes e pouco informativas.
5. **Matriz de coocorrência:** representação distribuída baseada na relação de proximidade entre palavras, permitindo capturar padrões mais contextuais.
6. **Word2Vec:** modelo de vetorização semântica treinado com gensim, gerando embeddings que capturam relações de similaridade e proximidade contextual entre palavras.
7. **Observações**

A variável sentimentScorePortugues, originalmente contínua, apresentava valores entre -1 e 1. Seguindo recomendações da própria API e práticas comuns na literatura, os escores foram particionados em três faixas para permitir a construção de uma variável categórica:



- **Negativo: valores entre -1 e -0.25**
- **Neutro: valores entre -0.25 e 0.25**
- **Positivo: valores entre 0.25 e 1**

Assim, definimos três classes explícitas: *negative*, *neutral* e *positive*. Essa discretização se mostrou adequada para modelagem supervisionada, permitindo treinar classificadores tradicionais e comparar o impacto das diferentes estratégias de pré-processamento e extração de atributos.

8. Resultados e discussão

Features	Pré-processado	Acurácia
Bag of Words	NÃO	0.833850
TF-IDF	NÃO	0.815685
Coocorrência	NÃO	0.853788
Word2Vec	NÃO	0.793088
Estatísticas	NÃO	0.793974
Bag of Words	SIM	0.826761
TF-IDF	SIM	0.816128
Coocorrência	SIM	0.850244
Word2Vec	SIM	0.794417
Estatísticas	SIM	0.793974

A comparação inicial entre os modelos sem pré-processamento e com pré-processamento mostrou uma diferença clara de desempenho. Sem qualquer tratamento dos textos, os modelos apresentaram maior dispersão do vocabulário, mais ruído linguístico e dificuldade em separar as três classes de sentimento. Com o pré-processamento completo — normalização, remoção de URLs e menções, stopwords, stemming e lematização — houve melhora consistente na acurácia, na estabilidade dos modelos e na separação entre sentimentos, principalmente entre positivo e neutro.



Ao analisar apenas os modelos **com pré-processamento**, as técnicas de extração de atributos apresentaram desempenhos distintos. As estatísticas individuais foram as menos eficazes, pois capturam pouca informação semântica. CountVectorizer apresentou resultados sólidos, mas o **TF-IDF** mostrou melhor desempenho geral, destacando termos relevantes e reduzindo o impacto de palavras muito frequentes. Word2Vec teve desempenho competitivo ao capturar relações semânticas, embora dependesse fortemente de um bom pré-processamento para evitar ruído. A matriz de coocorrência obteve resultados intermediários.

No conjunto, as melhores combinações envolveram TF-IDF com modelos lineares e Word2Vec com modelos não lineares, sendo o TF-IDF o mais estável e eficiente na classificação das três classes — negative, neutral e positive — definidas a partir da discretização do “sentimenteScorePortuguese”.

Pré-processamento	Acurácia	Feature Extraction
Lemmatização	0.850244	Coocorrência
Stemming	0.853345	Coocorrência

Os resultados mostram que, usando o extrator de coocorrência, tanto a lemmatização quanto o stemming apresentam acurácias muito próximas (0.850 e 0.853). Isso indica que a coocorrência é pouco sensível à forma exata das palavras, já que trabalha principalmente com relações de proximidade no texto. O leve ganho do stemming pode ser explicado pela maior redução do vocabulário, tornando a matriz mais compacta. No geral, ambos os métodos funcionam de forma semelhante com esse tipo de extração.

No geral, os resultados sugerem que:

9. O extrator de coocorrência é relativamente robusto à escolha entre stemming e lematização.
10. O stemming oferece leve vantagem em acurácia, possivelmente devido à maior redução da dimensionalidade.
11. Ambos os métodos são adequados para esse tipo de extração, com diferenças marginais de desempenho.



12. Conclusões

A análise realizada demonstra que o pré-processamento adequado exerce papel fundamental na melhoria da classificação de sentimentos em manchetes políticas. As etapas de normalização, remoção de ruídos, stopwords, stemming e lematização contribuíram para reduzir a variabilidade do texto e fortaleceram o desempenho dos modelos em relação aos experimentos sem pré-processamento.

Entre as técnicas de extração de atributos, o TF-IDF destacou-se como a abordagem mais consistente e eficiente, oferecendo melhor equilíbrio entre simplicidade, desempenho e estabilidade. O Word2Vec apresentou bons resultados ao capturar nuances semânticas mais profundas, embora dependa de maior cuidado na preparação dos dados. Já a coocorrência mostrou-se robusta a diferentes técnicas de redução lexical, apresentando resultados similares com lematização e stemming.

A discretização do *sentimentScorePortugues* em três classes — *negative*, *neutral* e *positive* — mostrou-se adequada para a tarefa, facilitando a separação entre categorias e permitindo que os modelos identificassem padrões editoriais presentes nas manchetes.

De modo geral, os resultados evidenciam que a combinação de um pipeline sólido de pré-processamento com técnicas apropriadas de vetorização é essencial para obter classificadores mais precisos no domínio político. Trabalhos futuros podem explorar modelos baseados em transformadores, análises temporais da cobertura jornalística e estudos mais aprofundados sobre possíveis vieses entre diferentes veículos de notícias.