

Classificando Aceitação de Artigos no Dataset PeerRead

Arthur Cisotto Machado
Engenharia da Computação
Insper - Instituto de Ensino e Pesquisa
São Paulo, Brasil
arthurcm5@al.insper.edu.br

Abstract—Este estudo aplica técnicas de processamento de linguagem natural e aprendizado de máquina ao dataset PeerRead para prever a aceitação de artigos científicos. Utilizando apenas o texto dos abstracts, implementamos um pipeline que inclui pré-processamento de texto, vetorização TF-IDF e classificação por regressão logística. O modelo alcançou uma acurácia balanceada de 0,6534, comparável aos resultados do estudo original. Analisamos o impacto do tamanho do dataset na performance do modelo e realizamos modelagem de tópicos para identificar áreas temáticas e suas respectivas taxas de aceitação. Nossos resultados demonstram a viabilidade de prever aceitação de artigos com precisão moderada usando métodos relativamente simples, ao mesmo tempo que oferecemos insights interpretáveis sobre os fatores que influenciam a aceitação. Este trabalho contribui para a compreensão do processo de revisão por pares e sugere direções para futuras pesquisas em análise automatizada de publicações científicas.

I. DATASET

Este estudo utiliza o dataset PeerRead [1], uma coleção de artigos científicos e suas revisões por pares. O objetivo principal é desenvolver um modelo preditivo para a aceitação ou rejeição de artigos com base em seu conteúdo e metadados associados.

A análise da distribuição de papers (Fig. 1) revela uma preponderância de artigos provenientes do arXiv, particularmente nas áreas de aprendizado de máquina (cs.lg) e inteligência artificial (cs.ai). Esta distribuição não uniforme pode influenciar a generalização do modelo, potencialmente favorecendo características mais comuns em preprints do que em artigos de conferências revisados por pares.

As taxas de aceitação (Fig. 2) apresentam variações significativas entre os datasets, refletindo potencialmente diferentes padrões de revisão e qualidade média dos artigos submetidos.

II. PIPELINE DE CLASSIFICAÇÃO

O pipeline de classificação implementado consiste em três etapas principais:

- 1) **Pré-processamento:** Inclui tokenização, remoção de stopwords, stemming e lematização, visando normalizar e limpar o texto dos abstracts.
- 2) **Vetorização:** Utiliza-se TF-IDF (Term Frequency-Inverse Document Frequency) com 5000 features para transformar o texto em representações numéricas.

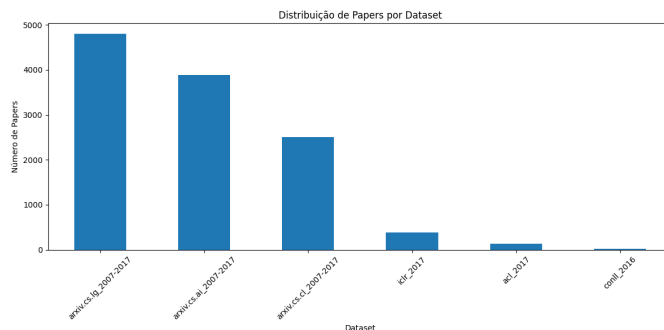


Fig. 1. Distribuição de papers por dataset no PeerRead, evidenciando a predominância de papers do arXiv em áreas como aprendizado de máquina e inteligência artificial.

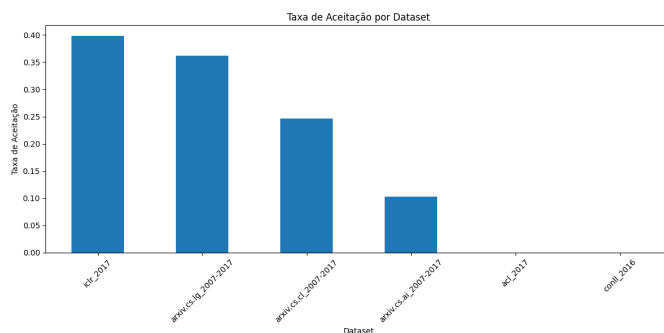


Fig. 2. Taxa de aceitação por dataset, destacando as diferenças nas taxas de aceitação entre conferências e repositórios de preprints.

- 3) **Classificação:** Emprega-se um modelo de Regressão Logística para a tarefa de classificação binária (aceito/rejeitado).

A abordagem bag-of-words foi escolhida devido à sua eficácia em capturar termos relevantes para a qualidade e aceitação de artigos científicos, embora reconheçamos suas limitações em capturar contexto e ordem das palavras.

III. AVALIAÇÃO

O modelo de camada única alcançou uma acurácia balanceada de 0,6534, alinhando-se com os resultados reportados no paper original [1]. A Fig. 3 ilustra a correlação entre a pontuação média dos revisores e a aceitação do artigo, com um coeficiente de 0,77, indicando uma forte relação positiva.

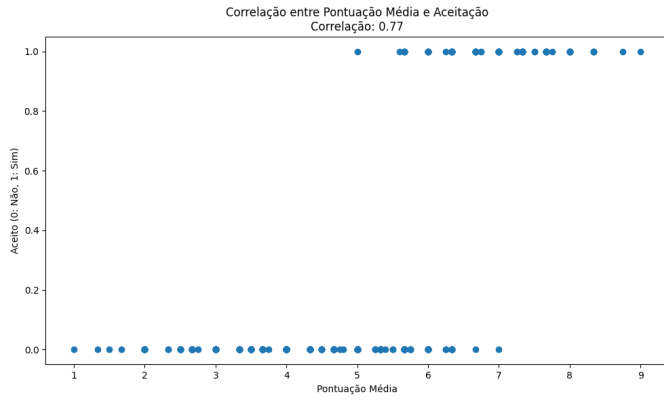


Fig. 3. Correlação entre pontuação média dos revisores e aceitação do artigo (coeficiente de correlação: 0,77).

IV. TAMANHO DO DATASET

A análise do efeito do tamanho do dataset na performance do modelo (Fig. 4) revela um aumento significativo na acurácia entre 10% e 25

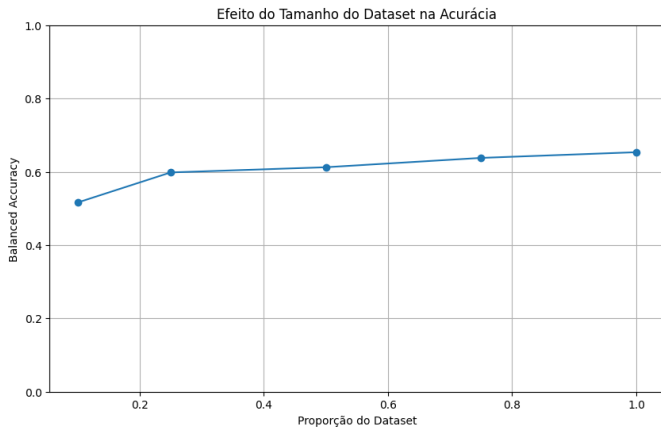


Fig. 4. Efeito do tamanho do dataset na acurácia do modelo, demonstrando ganhos significativos iniciais seguidos por melhorias incrementais.

A comparação entre erros de treino e teste (Fig. 5) indica um leve overfitting, sugerindo que técnicas de regularização poderiam ser benéficas.

V. ANÁLISE DE TÓPICOS

A modelagem de tópicos utilizando Latent Dirichlet Allocation (LDA) identificou cinco tópicos principais:

- 1) Modelos de Linguagem e Processamento de Texto (taxa de aceitação: 0,26)
- 2) Algoritmos e Otimização (taxa de aceitação: 0,21)
- 3) Aprendizado Profundo e Redes Neurais (taxa de aceitação: 0,19)
- 4) Computação e Sistemas de Informação (taxa de aceitação: 0,10)
- 5) Algoritmos de Aprendizado e Análise de Dados (taxa de aceitação: 0,32)

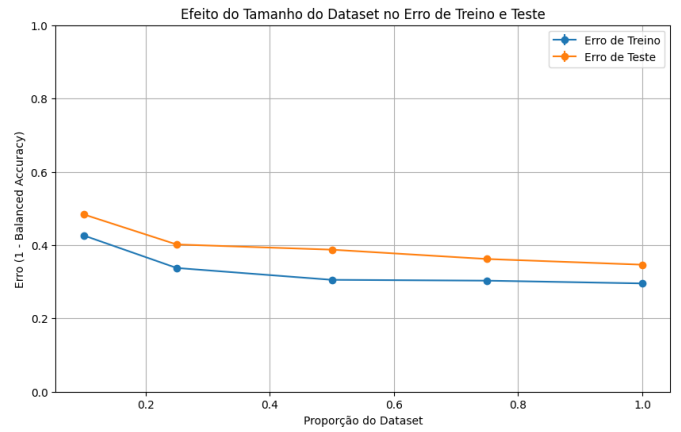


Fig. 5. Comparação entre erros de treino e teste em função do tamanho do dataset, sugerindo um leve overfitting do modelo.

Esta análise revelou variações nas taxas de aceitação entre diferentes áreas de pesquisa, com tópicos mais aplicados tendendo a ter taxas de aceitação ligeiramente mais altas.

VI. CONCLUSÃO E COMPARAÇÃO COM O PAPER ORIGINAL

Este estudo demonstra a viabilidade de prever a aceitação de artigos científicos com acurácia moderada utilizando apenas o texto do abstract. Nossa abordagem, embora mais simples, alcançou resultados comparáveis ao paper original [1], que reportou uma acurácia de aproximadamente 0,65 para seu melhor modelo.

Pontos-chave de comparação:

- **Acurácia:** Nossa acurácia balanceada de 0,6534 é consistente com o paper original.
- **Abordagem:** Utilizamos uma abordagem mais simples (TF-IDF e Regressão Logística) em comparação com os modelos mais complexos do paper original, como redes neurais profundas.
- **Recursos:** Nosso modelo utilizou apenas o texto do abstract, enquanto o paper original incorporou metadados adicionais, como informações dos autores e citações.
- **Interpretabilidade:** Nossa análise de importância de features e modelagem de tópicos oferece insights interpretáveis, complementando a abordagem mais focada em desempenho do paper original.

Limitações e direções futuras incluem a incorporação de metadados adicionais, exploração de técnicas de NLP mais avançadas como transformers pré-treinados, e o desenvolvimento de modelos específicos para diferentes áreas de pesquisa ou conferências.

REFERENCES

- [1] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications," arXiv preprint arXiv:1804.09635, 2018.