Treinando um Classificador de Notícias

Arthur Faria

Geraldo Melo

Gustavo Ventino

Lucas Marques

Pedro Ribeiro

Universidade Federal de Uberlândia

Bacharelado em Ciências da Computação - Mineração de Dados

Murilo Guimarães Carneiro

Introdução

Neste trabalho desenvolvemos e avaliamos um classificador automático de artigos jornalísticos, utilizando o News Category Dataset sob diferentes formas de agrupamentos dos dados.

A tarefa principal consiste em definir a categoria de uma notícia com base em apenas o seu conteúdo textual, para isso vamos analisar a Manchete(Headline) e o Lide(Short_Description) do artigo.

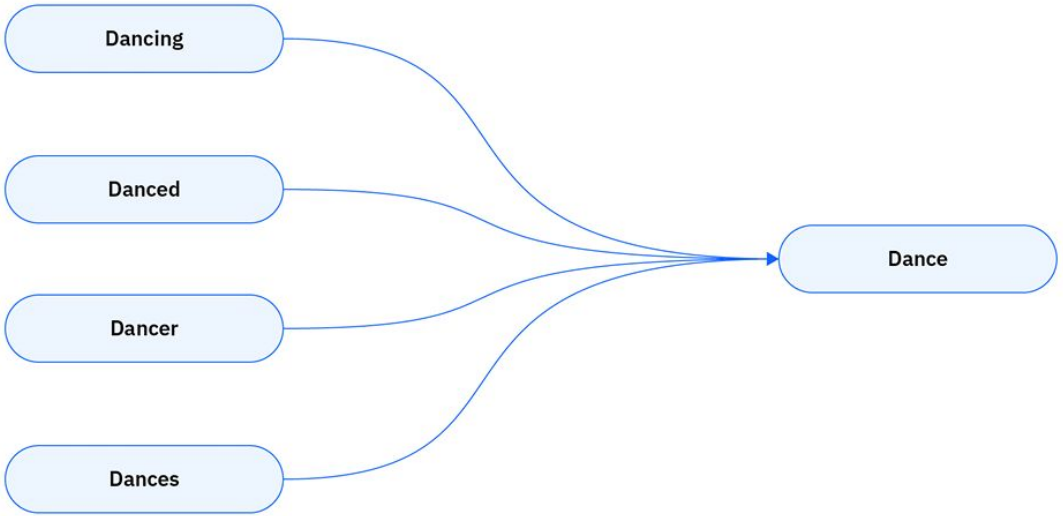
Metodologia

Nossa metodologia consiste das seguintes etapas e em suas respectivas ordens:

- Pré-Processamento: Remoção caracteres indesejados, ruídos e redução das palavras às suas formas canônicas, para facilitar a análise.
- Representação TF-IDF: Cada palavra recebe um peso proporcional à sua importância e frequência no texto.
- Classificação SVM: Busca encontrar o hiperplano que melhor separa as classes com a maior margem possível no espaço de características.
- Avaliação: Utilizamos a métrica de Acurácia Global, a qual mede a proporção de notícias corretamente classificadas em relação ao total.

Pré-Processamento

O pré-processamento textual incluiu a limpeza de dados com expressões regulares (remoção de URLs, menções e caracteres especiais), conversão para minúsculas, remoção de *stopwords* e lematização (mapear palavras para a sua “forma base”).



Classificador

As manchetes e lides foram transformadas em vetores numéricos usando a técnica **TF-IDF**, onde cada dimensão corresponde a um termo do vocabulário do corpus, e o valor nessa dimensão reflete a importância do termo.

Para isso, utilizamos o **Support Vector Machine (SVM)**,o qual busca o hiperplano ótimo que separa as classes com a maior margem possível. Juntamente com um **kernel linear**, adequado para espaços de alta dimensionalidade como os gerados por TF-IDF.

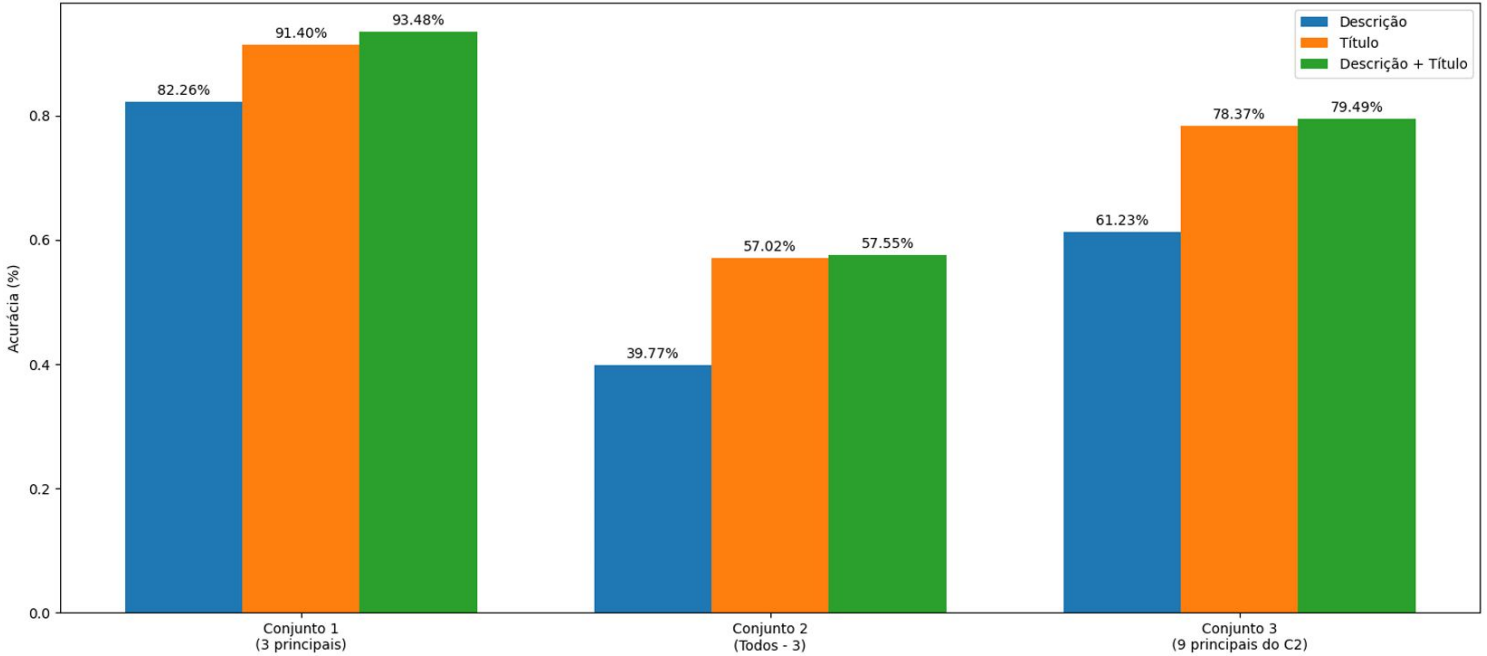
Experimentos

Para os experimentos, separamos o dataset em 3 grupos de categorias:

1. Categorias mais frequentes: Contém apenas as notícias das três categorias com maior número de instâncias no dataset.
2. Excluindo mais frequentes: Engloba todas as notícias que não estavam presentes entre as três mais frequentes.
3. 9 Categorias mais frequentes: Engloba as 9 categorias mais frequentes sem considerar o Top 3 analisado no primeiro agrupamento.

Resultados

Podemos perceber que conforme o número de categorias aumentam, a acurácia do classificador diminui consideravelmente.



Conclusão

Por fim, é possível concluir que o classificador baseado em TF-IDF + SVM linear é eficaz na categorização de notícias, especialmente em cenários com poucas classes e dados balanceados. Já que foi possível alcançar uma alta acurácia (>93%) no grupo com as categorias mais frequentes, utilizando a combinação de Título + Descrição.

Referências

[1] Misra, R. (2022) “News Category Dataset”, https://www.kaggle.com/datasets/rmisra/ news-category-dataset/data, Kaggle, acessado em: 4 de maio de 2025.

[2] Turing USP (2023) “Fernando Pessoa- Análise e Geração de Textos”, https://github.com/ turing-usp/fernando-pessoa, GitHub, acessado em: 4 de maio de 2025.

[3] Ribeiro, P. (2024) “GBC212- Repositório de Projetos”, https://github.com/PedroRibeiroA123/ GBC212, GitHub, acessado em: 4 de maio de 2025.

[4] IBM, "Stemming e lematização: qual a diferença?", IBM Think Brasil, disponível em: https: //www.ibm.com/br-pt/think/topics/stemming-lemmatization, acessado em: 4 de maio de 2025.

[5] Turing Talks, "Como Machine Learning consegue diferenciar heterônimosde Fernando Pessoa", *Medium*,disponível em: https://medium.com/turing-talks/como-machine-learning-consegue-diferenciar-heter%C3% B4nimos-de-fernando-pessoa-156d0d52a478, acessado em: 4 de maio de 2025.