Treinando um Classificador de Notícias

Faria. Arthur, Melo. Geraldo, Ventino. Gustavo, Ribeiro. Pedro, Tiago. Lucas
Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
arhur.faria@ufu.br , pedro.ribeiro1@ufu.br, geraldo.ne@ufu.br, gustavo.ventino01@ufu.br, lucas.mpt@ufu.br

Resumo. *Este trabalho prático desenvolve um classificador automático para notícias usando aprendizado de máquina. Os dados textuais (manchetes e descrições) foram pré-processados (limpeza, lematização) e as categorias codificadas numericamente. Devido à frequência desigual das categorias, o dataset foi segmentado em três grupos para análise. Utilizou-se TF-IDF para vetorização textual e modelos SVM com kernel linear para classificação. Os resultados mostraram alta acurácia (até 93.485%) na classificação dos 3 temas mais frequentes, principalmente combinando manchete e descrição. A classificação de um número maior de categorias, apresentou acurácia menor (até 79.490%), evidenciando os desafios com dados desbalanceados e mais classes.*

1. Introdução

Este trabalho prático visa desenvolver e avaliar um classificador automático para notícias utilizando o News Category Dataset em diferentes agrupamentos dos dados. O dataset compreende um extenso conjunto de artigos de notícias, cada um associado a uma categoria, política, saúde, entretenimento, entre outras.

A tarefa específica abordada é a de atribuir uma categoria a uma dada notícia com base em seu conteúdo textual, utilizando a manchete (headline) e o lide (short_description) do artigo.

A metodologia empregada abrange as etapas fundamentais de um projeto de aprendizado de máquina aplicado a texto. O pré-processamento dos dados textuais foi a etapa inicial e indispensável, focando na limpeza dos textos para remover caracteres indesejados e ruídos, e na lematização para reduzir as palavras às suas formas canônicas, minimizando a variabilidade e facilitando a identificação de padrões. Para a representação numérica do texto pré-processado, foi selecionada a técnica de vetorização TF-IDF (Term Frequency-Inverse Document Frequency), que pondera a importância de cada termo em um documento em relação à sua frequência no corpus inteiro.

Um aspecto crítico observado no dataset é a distribuição heterogênea da frequência das categorias, caracterizando um cenário de dados desbalanceados. Para investigar o impacto deste fenômeno no desempenho do classificador, optou-se por segmentar o dataset em diferentes agrupamentos de categorias, permitindo avaliar o modelo sob distintas condições de balanceamento e número de classes.

Para a tarefa de classificação propriamente dita, foi empregado o algoritmo Support Vector Machine (SVM) configurado com um kernel linear. O SVM é um modelo discriminativo robusto, particularmente eficaz em espaços de alta dimensionalidade como os gerados por representações TF-IDF. Ele busca encontrar o hiperplano ótimo que maximiza a margem de separação entre as classes no espaço de características. O kernel linear assume uma separabilidade linear (ou aproximadamente linear) dos dados neste espaço, uma suposição frequentemente válida e computacionalmente eficiente para problemas de classificação de texto. A avaliação do desempenho do modelo foi realizada na etapa de pós-processamento, utilizando a métrica de acurácia global, que quantifica a proporção de instâncias corretamente classificadas pelo modelo em relação ao total.

2. Desenvolvimento

A fase de desenvolvimento deste trabalho prático foi dedicada à implementação e avaliação do pipeline de classificação automática de notícias, com atenção especial aos desafios apresentados pela distribuição desigual das classes.

O pré-processamento textual foi implementado através de uma sequência programática de transformações. A limpeza envolveu a aplicação de expressões regulares para identificar e remover padrões como URLs, menções e caracteres especiais. A conversão para minúsculas uniformizou a capitalização dos termos. A remoção de stopwords utilizou uma lista predefinida de palavras funcionalmente irrelevantes. A lematização foi aplicada mapeando as palavras às suas formas base como demonstrado na figura 1. Por exemplo, formas flexionadas do verbo "correr"em português seriam reduzidas ao seu lema: "corria", "correrão", "correram", "correria"→ "correr". Este passo foi fundamental para agrupar semanticamente termos com variações morfológicas. A combinação do texto da manchete e da descrição curta foi realizada concatenando-os em um único campo textual após o pré-processamento individual de cada um, criando uma representação textual unificada por notícia.

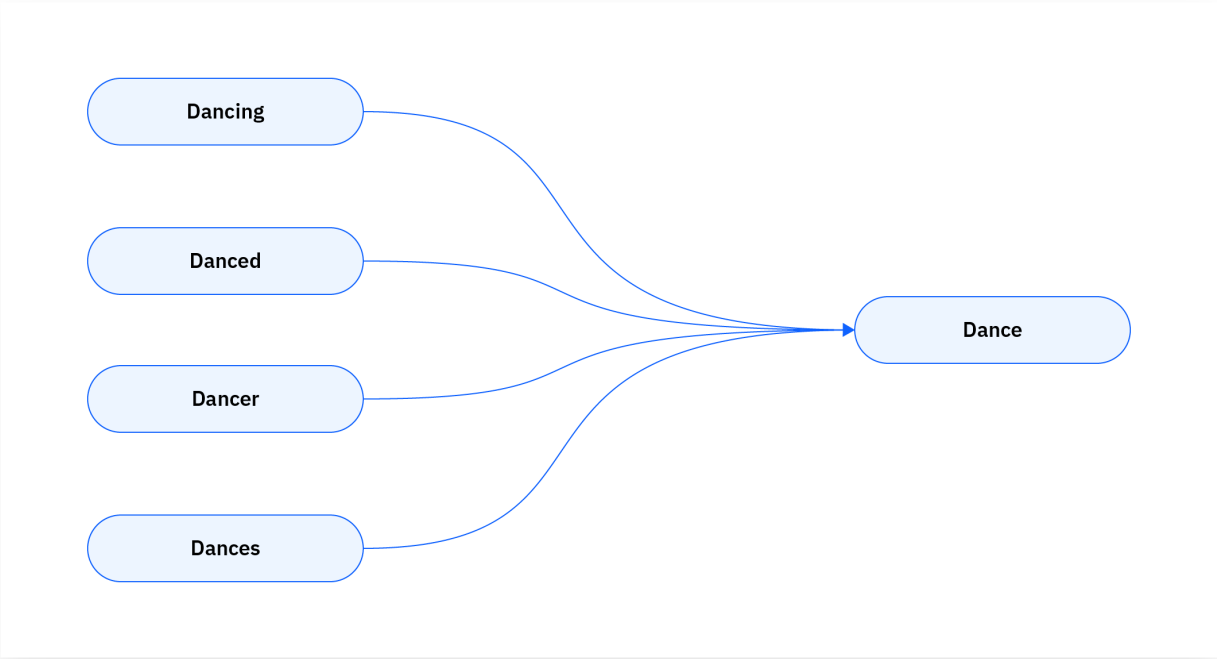

Figura 1: Exemplo de lematização com palavras derivadas da palavra "Dance" [4]

Aplicou-se a vetorização utilizando TF-IDF sobre o texto combinado. Esta técnica gera vetores numéricos onde cada dimensão corresponde a um termo do vocabulário do corpus, e o valor nessa dimensão reflete a importância do termo. Palavras que aparecem frequentemente em uma notícia específica mas são raras em outras, recebem um peso maior, tornando-se importantes para diferenciar categorias. Após a aplicação dessas tecnicas, obtivemos uma base de dados como demonstrada na figura 2

A estratégia de segmentação foi separar os datasets em 3 grupos de categorias:

1. Grupo das 3 Categorias Mais Frequentes: Contendo apenas as notícias pertencentes às três categorias com maior número de instâncias no dataset original (empiricamente identificadas como

	headline	category	short_description	headline_processado	desc_processado	texto_tot
0	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	million americans roll sleeves omicrontargeted...	health experts say early predict whether deman...	million americans roll sleeves omicrontargeted...
1	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	american airlines flyer charge ban life punch ...	subdue passengers crew flee back aircraft conf...	american airlines flyer charge ban life punch ...
2	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	funniest tweet cat dog week sept	dog dont understand could eat	funniest tweet cat dog week septdog dont under...
3	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	funniest tweet parent week sept	accidentally put grownup toothpaste toddlers t...	funniest tweet parent week septaccidentally pu...
4	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	woman call cop black birdwatcher lose lawsuit ...	amy cooper accuse investment firm franklin tem...	woman call cop black birdwatcher lose lawsuit ...

Figura 2: Base de dados após o pré-processamento

POLITICS, WELLNESS, ENTERTAINMENT). Este grupo representa um cenário de classificação multiclasse com baixo número de classes e relativa proximidade de balanceamento entre elas.

2. Grupo Excluindo as 3 Categorias Mais Frequentes: Englobando todas as notícias cujas categorias não estavam entre as três mais frequentes. Este grupo caracteriza um cenário com um número maior de classes e, potencialmente, um grau mais elevado de desbalanceamento, pois inclui diversas categorias minoritárias.

3. Grupo das 9 Categorias Mais Frequentes (Excluindo as 3 Principais): Constituído pelas notícias das nove categorias mais frequentes, porém excluindo o top 3 já analisado separadamente. Este grupo representa um cenário intermediário em termos de número de classes e desbalanceamento, permitindo observar o comportamento do modelo em uma condição de complexidade moderada.

Esses diferentes grupos foram todos treinados usando Support Vector Machine (SVM) com kernel linear. O SVM busca encontrar o hiperplano (como demonstrado na figura 3) ótimo que melhor separa as classes no espaço de características definido pelos vetores TF-IDF, maximizando a margem entre as classes. O kernel linear é frequentemente eficaz para dados textuais devido à sua capacidade de lidar com a alta dimensionalidade e por assumir que as classes são linearmente separáveis ou aproximadamente separáveis no espaço de features.

Os resultados obtidos demonstraram claramente a influência do número e da distribuição das classes no desempenho da classificação. Para o grupo das 3 categorias mais frequentes (POLITICS, WELLNESS, ENTERTAINMENT), a acurácia alcançada foi notavelmente alta, atingindo até 93.485% como pode ser visto na figura 4. Porém, a performance decaiu ao classificar grupos com número maior de categorias e desbalanceamento com acurácia máxima atingindo 79.490%, como demonstrado na comparação dos testes na figura 5. O Naive-Bayes não foi utilizado por limitação de memória como pode ser visto na figura 6.

Os resultados obtidos para as classes majoritárias se alinham com o estado da arte para abordagens usando TF-IDF e SVM em datasets de notícias com características semelhantes. A queda de desempenho em cenários com desbalanceamento é um padrão consistente em estudos na área, refor-

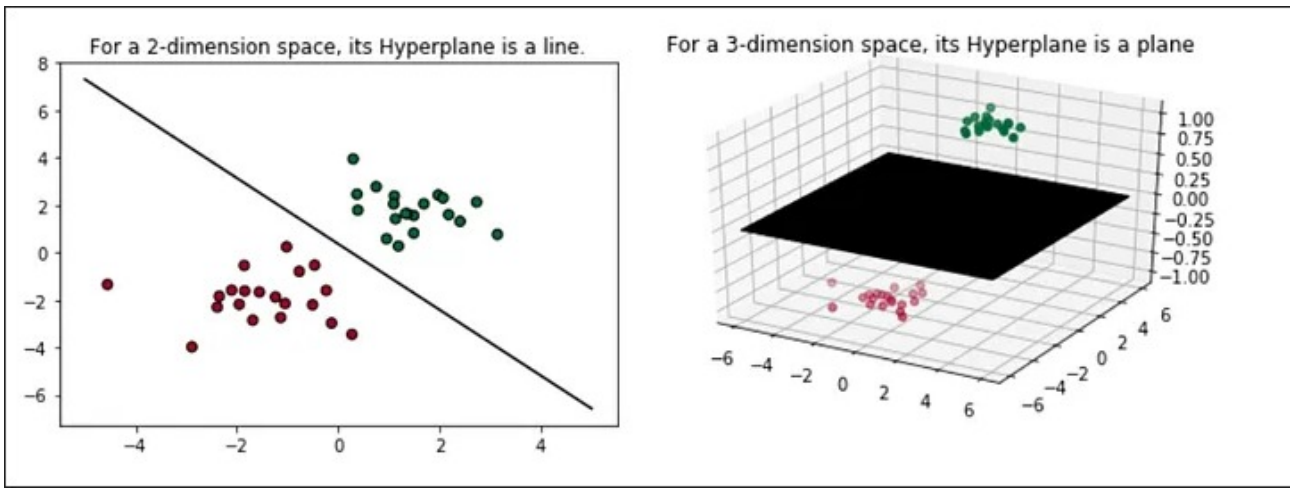

Figura 3: À direita demonstra o hiperplano em duas dimensões, À esquerda demonstra um espaço de 3 dimensões.

çando a necessidade de técnicas capazes de mitigar o problema.

3. Conclusões

Este trabalho prático explorou o desenvolvimento de um classificador automático para notícias utilizando o News Category Dataset e uma abordagem baseada em técnicas tradicionais de aprendizado de máquina. Os resultados obtidos confirmaram a eficácia do pipeline composto por pré-processamento textual (limpeza e lematização), vetorização com TF-IDF e classificação com SVM linear para a tarefa de categorização de notícias, particularmente em cenários com um número limitado de classes e distribuição de dados relativamente equilibrada.

O principal achado do trabalho reside na demonstração da viabilidade da classificação de notícias com alta acurácia (acima de 93%) para as categorias mais frequentes do dataset, evidenciando o potencial da abordagem para sumarizar e organizar grandes volumes de notícias em temas principais. A contribuição conjunta da manchete e da descrição para o aumento do desempenho nesta situação sublinha a importância de considerar múltiplos campos textuais quando disponíveis.

Entretanto, a análise do desempenho em grupos com maior número de categorias e acentuado desbalanceamento revelou as limitações inerentes à abordagem adotada em face desses desafios. A queda na acurácia para esses cenários (abaixo de 80%) reforça a necessidade de estratégias mais sofisticadas para lidar com a complexidade de diferenciar entre um número maior de classes e garantir um desempenho satisfatório para as categorias menos representadas.

Com base nos resultados e desafios identificados, os caminhos para pesquisas futuras são claros e promissoras. A aplicação de técnicas avançadas para tratamento de dados desbalanceados é um passo natural para melhorar a performance nas classes minoritárias. A exploração de modelos de representação de texto de última geração, como embeddings contextuais derivados de modelos de linguagem pré-treinados (e.g., BERT, RoBERTa), tem o potencial de capturar nuances semânticas que podem ser cruciais. A avaliação mais completa utilizando métricas como Precisão, Recall e F1-score, especialmente com foco no desempenho por classe, ofereceria uma visão mais precisa das áreas que necessitam de melhoria. Finalmente, a investigação de algoritmos de classificação alternativos, incluindo modelos de aprendizado profundo, pode levar a ganhos significativos de desempenho em datasets de larga escala e com alta complexidade de classes.

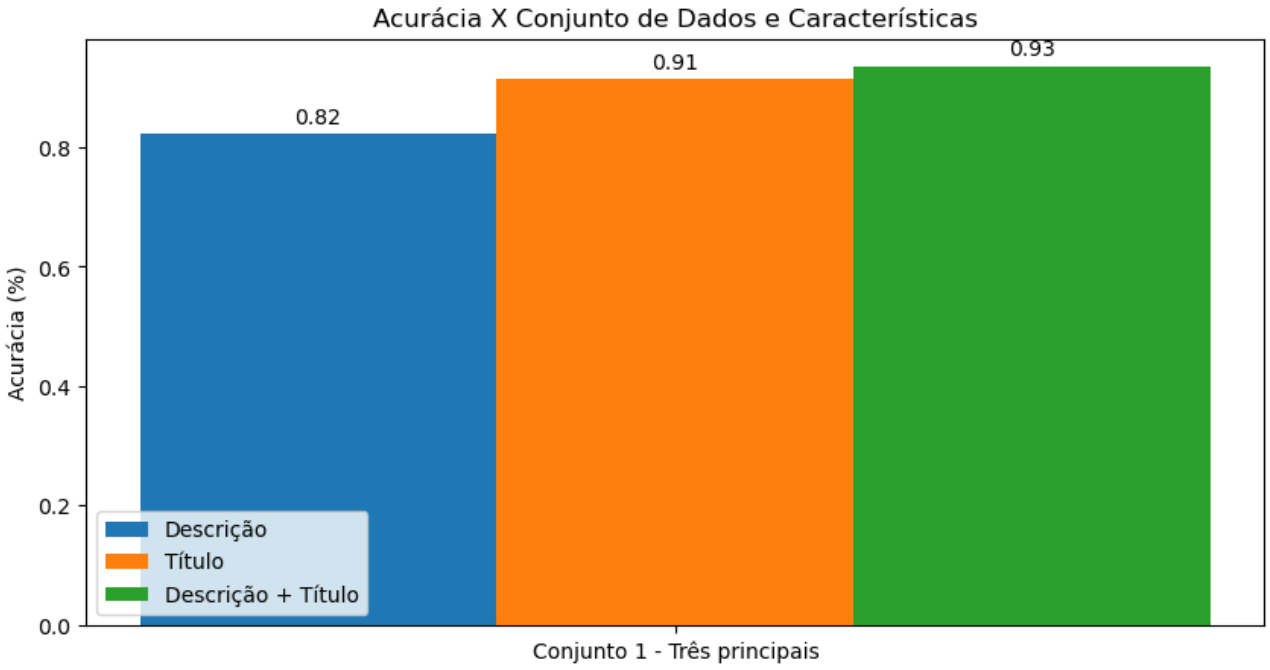

Figura 4: Resultados obtidos para o grupo das 3 categorias mais frequentes (POLITICS, WELLNESS, ENTERTAINMENT)

Referências

[1] Misra, R. (2022) "News Category Dataset", https://www.kaggle.com/datasets/rmisra/news-category-dataset/data, Kaggle. Acesso em 04 de maio de 2025.

[2] Turing USP (2023) "Fernando Pessoa - Análise e Geração de Textos", https://github.com/turing-usp/fernando-pessoa, GitHub. Acesso em 04 de maio de 2025.

[3] Ribeiro, P. (2024) "GBC212 - Repositório de Projetos", https://github.com/PedroRibeiroA123/GBC212, GitHub. Acesso em 04 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, Acesso em 04 de maio de 2025.

[5] Turing Talks, "Como Machine Learning consegue diferenciar heterônimos de 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.

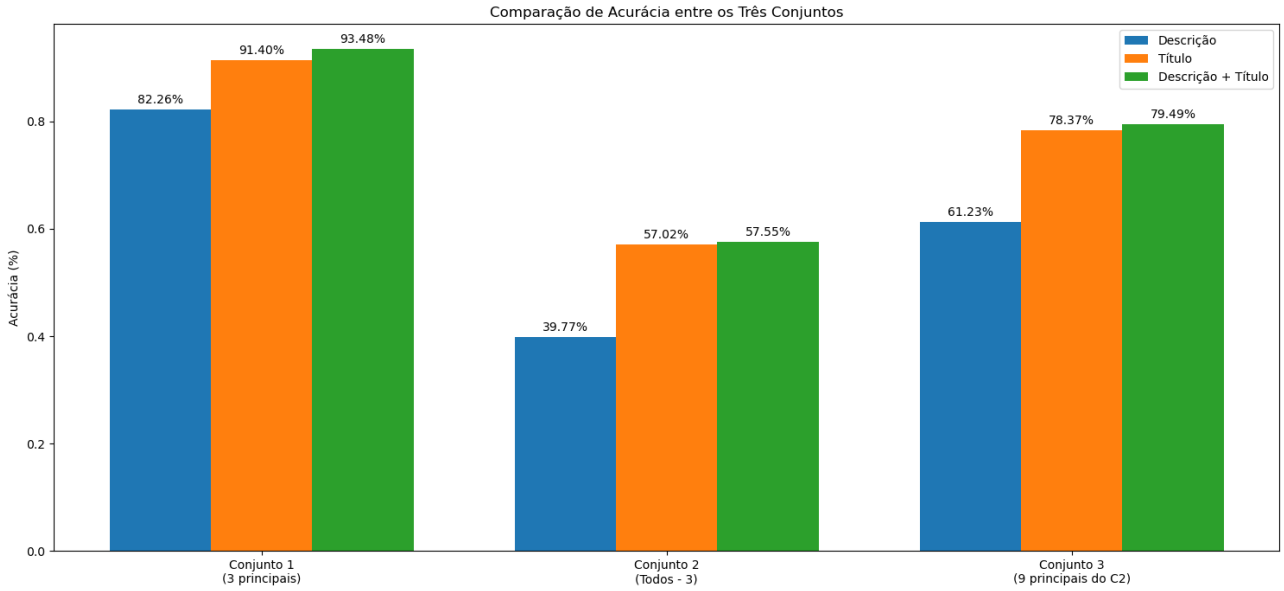

Figura 5: Comparações entre os conjuntos processados

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
Cell In[30], line 2
      1 nb = GaussianNB()
----> 2 nb.fit(train_X_tfidf_1desc.toarray(), train_Y1)
      3 predictions_nb = nb.predict(test_X_tfidf_1desc.toarray())
      4 acuracia = accuracy_score(predictions_nb, test_Y1)

File ~\miniconda3\Lib\site-packages\scipy\sparse\_compressed.py:1170, in _cs_matrix.toarray(self, order, out)
   1168 if out is None and order is None:
   1169     order = self._swap('cf')[0]
-> 1170 out = self._process_toarray_args(order, out)
   1171 if not (out.flags.c_contiguous or out.flags.f_contiguous):
   1172     raise ValueError('Output array must be C or F contiguous')

File ~\miniconda3\Lib\site-packages\scipy\sparse\_base.py:1367, in _spbase._process_toarray_args(self, order, out)
   1365     return out
   1366 else:
-> 1367     return np.zeros(self.shape, dtype=self.dtype, order=order)

MemoryError: Unable to allocate 18.5 GiB for an array with shape (63818, 38933) and data type float64
```

Figura 6: Erro obtido ao utilizar o método Naive-Bayes