BERTimbau News – Um Refinamento do Modelo Voltado à Classificação de Notícias

Pedro Parentoni de Almeida Tiago Gomes da Silva

Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brasil
pedro.parentoni@aluno.ufop.edu.br, tiago.gs@aluno.ufop.edu.br

Abstract

Abstract (English): This paper presents a research project developed in the Neural Networks course (BCC406), proposing the implementation of a model for news categorization using BERTimbau as the basis for text analysis and classification. The results demonstrate a significant improvement in the performance of the trained model compared to the base model.

Resumo (Português): Este artigo apresenta um projeto de pesquisa desenvolvido na disciplina de Redes Neurais (BCC406), turma 24.2, propondo a implementação de um modelo para categorização de notícias utilizando o BERTimbau como base para análise e classificação textual. Os resultados demonstram uma melhoria expressiva no desempenho do modelo treinado quando comparado ao modelo base.

1 Introdução

O presente estudo propõe a utilização do modelo BERTimbau [1] para a classificação de notícias em língua portuguesa. O objetivo é desenvolver um pipeline capaz de classificar notícias automaticamente em categorias principais.

A utilização do BERT [2] revolucionou o campo do processamento de linguagem natural (PLN), proporcionando resultados superiores em tarefas de classificação textual. Nesse contexto, o BERTimbau surge como uma adaptação para o português brasileiro, treinado em grandes corpora nacionais, garantindo um desempenho mais adequado ao idioma.

2 Trabalhos Relacionados

Nos últimos anos, diversas abordagens têm sido propostas para a classificação de textos, incluindo métodos clássicos e técnicas baseadas em redes neurais profundas. Dentre essas abordagens, destaca-se o modelo BERT [2] por sua capacidade de capturar informações contextuais de forma eficaz. Para a língua portuguesa, o BERTimbau [1] tem sido utilizado com sucesso, demonstrando avanços significativos na compreensão e classificação textual.

3 Metodologia

A metodologia proposta consiste nas seguintes etapas:

3.1 Coleta de Dados

Utilizou-se o dataset *News Articles in Brazilian Portuguese* [3], que contém notícias distribuídas em diferentes categorias.

3.2 Treinamento do Modelo

Empregou-se o BERTimbau [1] para a classificação das notícias em cinco categorias principais, listadas na Tabela 1.

Table 1: Categorias utilizadas na classificação de notícias

Categoria	**Descrição**
1	Ciência e Tecnologia
2	Economia
3	Entretenimento
4	Esportes
5	Política

O dataset contém cerca de 500 notícias em português. O modelo foi treinado por 20 épocas, com `batch_size` igual a 8, e posteriormente salvo como `modelo_treinado.pth`. O treinamento foi realizado utilizando 80% do dataset para treino e 20% para teste, procedimento que também foi mantido na comparação com o modelo base.

3.3 Configurações do Ambiente

Table 2: Especificações do Ambiente de Treinamento

Componente	**Especificação**
Memória RAM	32GB
SSD NVMe 1	Gen 4, 7GB/s
SSD NVMe 2	Gen 3, 3GB/s
SSD SATA	480GB, 500MB/s
Processador	Xeon E5-2680v4 (14 núcleos / 28 threads, 3.3GHz)
GPU	RTX 3060, 12GB
HD	1TB, 100MB/s
Fonte	1050W reais

4 Experimentos e Resultados

4.1 Resultados do Treinamento

O treinamento do modelo, utilizando o dataset *News Articles in Brazilian Portuguese* [3], alcançou uma acurácia de 95%. A Tabela 3 apresenta a perda por época durante o processo de treino:

Table 3: Perda por época durante o treinamento

Época	**Perda**
1	26.7485
2	7.9625
3	4.5731
4	2.6816
5	1.6875
6	1.6846
7	2.4687
8	0.9319
9	0.8010
10	0.8014
11	0.4417
12	0.3028
13	0.2971
14	0.3039
15	0.2161
16	0.2402
17	0.2558
18	0.2202
19	0.3224
20	0.2953

4.2 Comparação entre Modelo Base e Modelo Treinado

Realizou-se uma comparação entre o modelo BERTimbau base e o modelo treinado. O modelo base apresentou uma precisão de 27%, enquanto o modelo treinado alcançou 95%.

4.2.1 Matrizes de Confusão

Modelo Base Nos dados de teste, o modelo base obteve precisão de 27%. A Figura 1 apresenta a matriz de confusão correspondente.

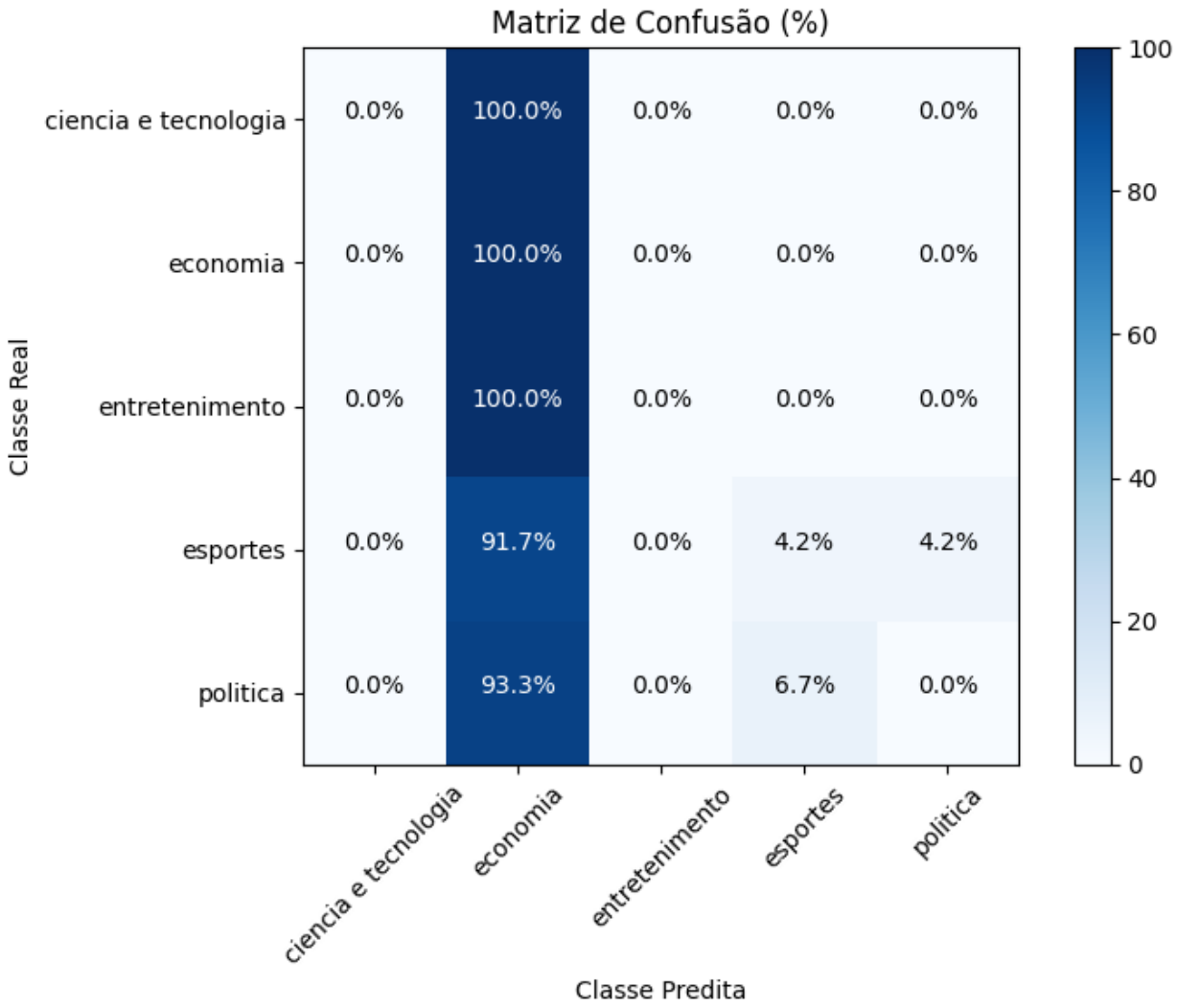

Figure 1: Matriz de confusão do Modelo BERTimbau base nos dados de treino.

Modelo Treinado O modelo treinado apresentou precisão de 95% nos dados de teste. A Figura 2 mostra a matriz de confusão correspondente.

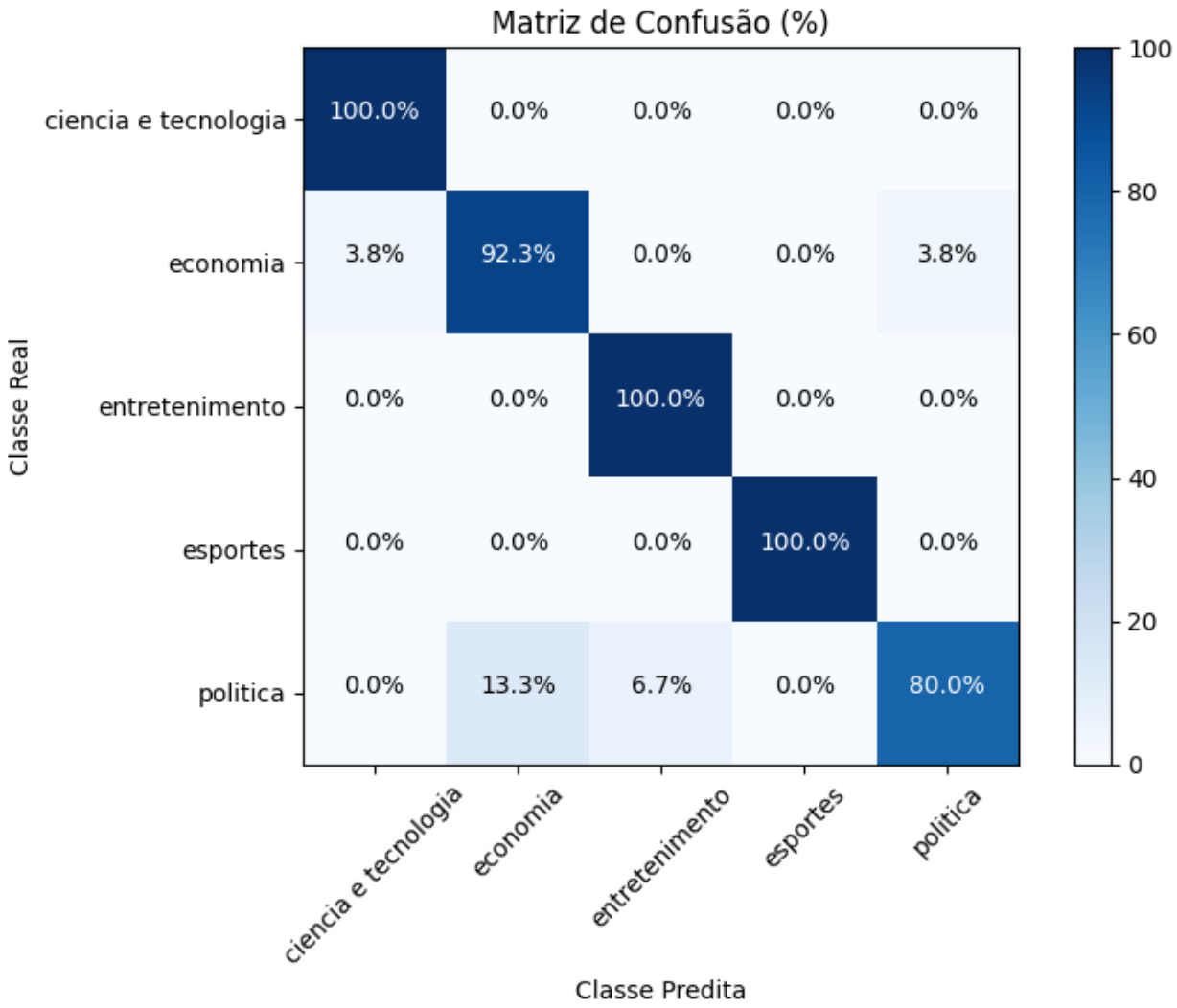

Figure 2: Matriz de confusão do Modelo BERTimbau treinado nos dados de treino.

4.3 Teste em Dataset Desconhecido

Para avaliar a robustez do modelo, testou-se sua aplicação em um dataset diferente, o *Notícias publicadas no Brasil* [4]. Para essa comparação, utilizamos 90% do segundo dataset.

Modelo Base No novo dataset, o modelo base obteve precisão de 20.05%. A Figura 3 apresenta a matriz de confusão correspondente.

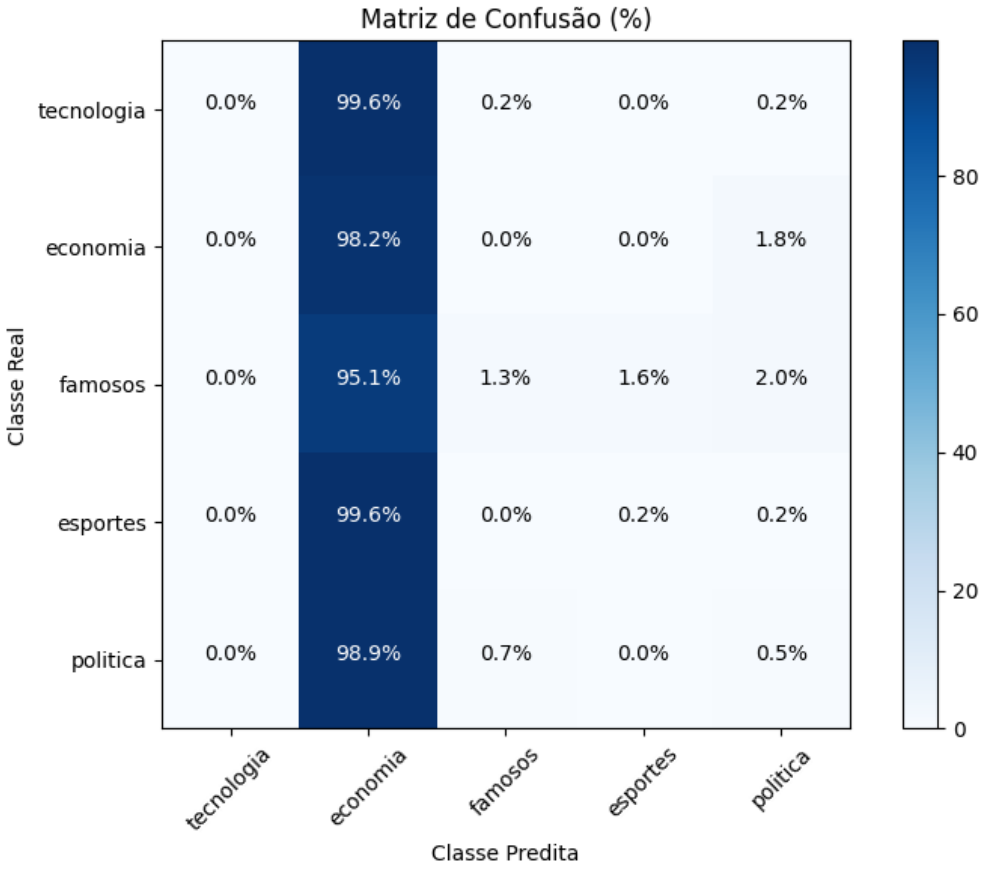

Figure 3: Matriz de confusão do Modelo BERTimbau base nos dados do dataset *Notícias publicadas no Brasil*.

Modelo Treinado No mesmo dataset, o modelo treinado alcançou precisão de 87.15%. A Figura 4 ilustra a matriz de confusão correspondente.

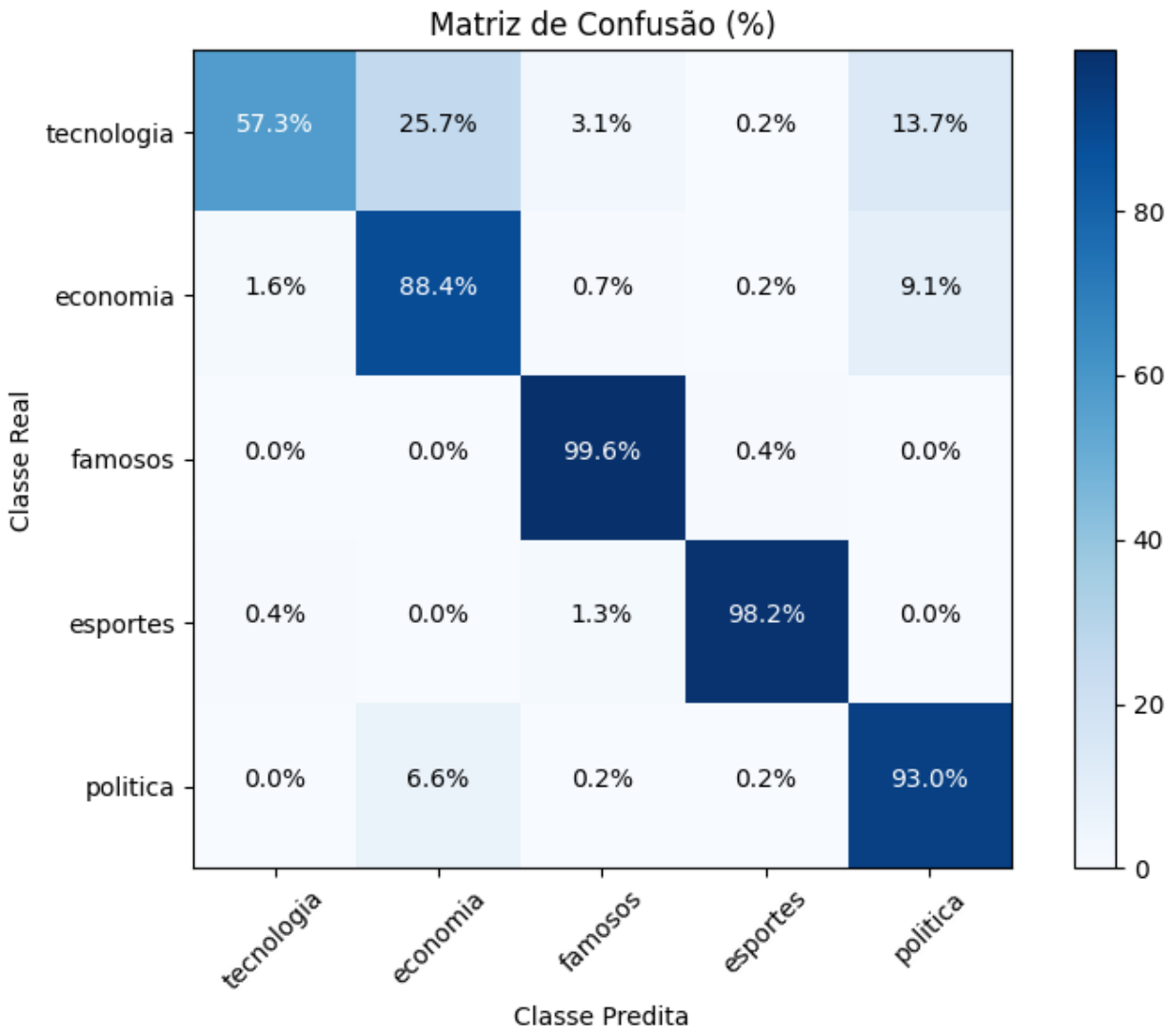

Figure 4: Matriz de confusão do Modelo BERTimbau treinado nos dados do dataset *Notícias publicadas no Brasil*.

5 Discussão

Os resultados demonstram que o treinamento do modelo BERTimbau levou a uma melhoria significativa no desempenho, com acurácia de 95% em comparação aos 27% do modelo base. Durante os testes em um dataset não visto, observou-se que:

- O modelo treinado lidou muito bem com uma nova classe, "famosos", ausente nas cinco categorias originais. A acurácia para essa label atingiu 99.6%.
- O modelo base apresentou uma tendência a classificar a maioria das notícias como "economia".

- Houve confusão entre as classes de política, tecnologia e economia no modelo treinado, possivelmente devido à interligação dos temas ou à ambiguidade nas categorizações presentes nos datasets.

Vale ressaltar que o procedimento de treinamento utilizou 80% do dataset para treino e 20% para teste, enquanto a comparação no segundo dataset foi realizada com 90% dos dados.

6 Conclusões

Este trabalho demonstrou que o treinamento do modelo BERTimbau para a classificação de notícias resulta em uma melhoria expressiva de desempenho em relação ao modelo base. Os experimentos evidenciaram não só a robustez do modelo treinado em contextos conhecidos, mas também sua capacidade de generalizar para dados não vistos, mesmo com classes não previstas originalmente.

References

[1] Souza, et al. BERTimbau: Pretrained BERT Models for Brazilian Portuguese.

[2] Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[3] Seiji, et al. News Articles in Brazilian Portuguese Dataset.

[4] Diogo, et al. Notícias publicadas no Brasil Dataset.