Classificação de Trocadilhos em Português com BERTimbau Large: Desafios e Resultados

Êmylle Beatriz de Sousa, Aislan Rafael R. de Sousa, Rafael T. Anchiêta

[1]Instituto Federal do Piauí – Picos – (IFPI)

emyllebeatrizdesousa@gmail.com, {aislanrafael,rta}@ifpi.edu.br

Abstract. *This work presents a study on the automatic detection of puns in Portuguese. For this purpose, the Puntuguese corpus was used, an unprecedented collection of 3,990 short texts, evenly divided between puns and non-puns. The BERTimbau Large model was applied to analyze the sentences, as it is specialized in the variations of Brazilian and European Portuguese. The results were satisfactory, even in light of the difficulty that the task imposes, showing that the model is promising for identifying this type of verbal humor. The study contributes to the field of Natural Language Processing in Portuguese and paves the way for future improvements, such as the addition of contextual information and more effective regularization techniques.*

Resumo. *Este trabalho apresenta um estudo sobre a detecção automática de trocadilhos em português. Para isso, foi utilizado o corpus Puntuguese, uma coleção inédita de 3.990 textos curtos, divididos igualmente entre trocadilhos e não trocadilhos. O modelo BERTimbau Large foi aplicado para analisar as sentenças, por ser especializado nas variações do português brasileiro e europeu. Os resultados foram satisfatórios, mesmo diante da dificuldade que a tarefa impõe, mostrando que o modelo é promissor para identificar esse tipo de humor verbal. O estudo contribui com a área de Processamento de Linguagem Natural em português e abre espaço para melhorias futuras, como a adição de informações de contexto e técnicas de regularização mais eficazes.*

1. Introdução

O humor é uma das formas mais complexas e sutis de comunicação humana, desempenhando um papel fundamental na cultura, nas interações sociais e na produção de conteúdo digital. Em um mundo cada vez mais digitalizado, o reconhecimento automático de conteúdo humorístico tem se mostrado um desafio significativo para o campo do Processamento de Linguagem Natural (PLN), devido à subjetividade, ambiguidade e à diversidade de estilos e formas humorísticas. Dentro desse contexto, a tarefa de detecção de trocadilhos se destaca como um exemplo particular de humor baseado em jogos de palavras, ambiguidade semântica e sobreposição fonológica. Detectar trocadilhos automaticamente requer que o sistema compreenda nuances linguísticas e contextuais que muitas vezes escapam a abordagens tradicionais de PLN.

A motivação para a detecção automática de trocadilhos se fundamenta na crescente demanda por sistemas capazes de compreender e interagir de maneira mais natural com o conteúdo textual gerado em redes sociais, memes, *chats* e outras plataformas digitais. Em diversas aplicações, como moderação de conteúdo, geração de texto criativo,

assistentes virtuais e análise de sentimentos, a identificação precisa de elementos humorísticos — como os trocadilhos — pode melhorar significativamente a qualidade da interação e a precisão do sistema. No entanto, o reconhecimento de trocadilhos apresenta desafios consideráveis: sua natureza pode ser sutil, muitas vezes dependendo de duplo sentido ou conhecimento prévio, e as variações regionais e contextuais na língua portuguesa podem tornar o processo ainda mais complexo.

Para resolver esse problema, este trabalho apresenta um sistema de classificação de textos voltado para a detecção automática de trocadilhos, utilizando o modelo de linguagem BERTimbau large [Souza, Nogueira e Lotufo 2020]. Esse modelo, treinado no corpus Puntuguese [Inacio et al. 2024], especializado em português, se mostrou eficiente em várias tarefas de PLN, como a identificação de discurso de ódio, similaridade textual e reconhecimento de entidades nomeadas. Além disso,que permitiram a captura de padrões linguísticos específicos, como jogos de palavras e ambiguidades, que não são detectáveis apenas pela tokenização. Essas técnicas ajudaram a tornar o sistema mais robusto e preciso, permitindo identificar melhor os trocadilhos presentes nas sentenças.

Os principais resultados indicam uma melhoria significativa nos indicadores de avaliação da tarefa. O modelo alcançou uma acurácia de 70%. Esse resultado é competitivo e supera trabalhos anteriores, como o de [Inacio et al. 2024], que obteve uma acurácia de 68% com o corpus Puntuguese — também utilizado neste trabalho. Ainda assim, há espaço para melhorias em relação aos melhores desempenhos reportados na literatura. Embora a tarefa de detecção de trocadilhos seja desafiadora, os resultados obtidos por esse estudo representam um avanço importante na detecção de humor verbal. Considerando a complexidade do corpus de trocadilhos, os resultados obtidos neste trabalho representam uma contribuição significativa para a área de PLN com foco em humor verbal.

A estrutura do artigo está organizada da seguinte forma: na Seção 2, revisamos os principais trabalhos relacionados à detecção de humor e trocadilhos, na Seção 3, descrevemos em detalhes o *corpus Puntuguese*, suas características e origens, na Seção 4, apresentamos a metodologia adotada e os ajustes realizados no modelo BERTimbau Large; na Seção 5, discutimos os resultados experimentais, suas implicações e limitações; por fim, na Seção 6, apresentamos as conclusões e apontamos direções promissoras para pesquisas futuras.

2. Trabalhos Relacionados

A detecção de humor e, mais especificamente, a identificação de trocadilhos em textos é uma tarefa desafiadora no campo do Processamento de Linguagem Natural (PLN). A ambiguidade linguística, a dependência de contextos culturais e a natureza subjetiva do humor tornam essa tarefa particularmente difícil para os sistemas automáticos. Diversas abordagens têm sido exploradas para lidar com esses desafios, com modelos baseados em aprendizado de máquina e redes neurais se destacando como alternativas eficazes.

O trabalho de [Miller, Hempelmann e Gurevych 2017] propôs uma competição focada na detecção e interpretação de trocadilhos em inglês. Este estudo representa um avanço significativo na área, fornecendo uma avaliação ampla sobre diferentes abordagens para detectar trocadilhos e destacando a importância do contexto e das características semânticas para essa tarefa. Embora o estudo se concentre no inglês, ele fornece metodo-

logias e desafios valiosos que podem ser adaptados à detecção de trocadilhos em outros idiomas, como o português, embora seja necessário ajustar as abordagens para as particularidades linguísticas e culturais de cada língua.

Em uma linha de pesquisa semelhante sobre a detecção automática de humor, [Bonet, Rincón e López 2023] investigaram a identificação de humor prejudicial em textos do Twitter, utilizando dados em espanhol. Embora o estudo não se concentre em trocadilhos, ele aborda desafios similares relacionados às nuances semânticas e contextuais do humor textual. Os autores obtiveram resultados expressivos, com destaque para um F1-Macro de 0,895, evidenciando o potencial dos modelos baseados em transformadores. Ainda assim, o trabalho aponta limitações relacionadas ao *overfitting*, especialmente em contextos com conjuntos de dados pequenos — um problema também recorrente na detecção de trocadilhos.

Por outro lado, [Cruz et al. 2023] exploraram o uso de ensembles de modelos transformadores para detectar humor prejudicial e estereótipos em *tweets* escritos em espanhol, no contexto da competição HUHU@IberLEF 2023. A abordagem propôs um sistema de votação ponderada entre modelos como BERT, RoBERTa e BETO, combinando as previsões com base no desempenho individual de cada modelo. O sistema alcançou F1-macro de 79,6% na tarefa de classificação multilabel dos grupos-alvo, obtendo o 1º lugar entre 49 equipes participantes. Essa estratégia de combinar múltiplos modelos mostra-se promissora para a detecção de trocadilhos, por permitir capturar diferentes nuances linguísticas e contextuais, características centrais nesse tipo de humor.

O Puntuguese [Inacio et al. 2024] se destaca como um avanço relevante na detecção de trocadilhos em português. Desenvolvido para resolver limitações de corpora anteriores, como o utilizado por Gonçalo Oliveira et al. (2020), o Puntuguese passou por curadoria manual para evitar que os modelos aprendessem padrões superficiais, o que prejudicava sua capacidade de generalização — ou seja, de acertar exemplos novos de forma consistente. Com textos extraídos de fontes populares — *blogs*, Instagram e YouTube — e organização balanceada entre exemplos com e sem trocadilhos, o corpus se consolidou como uma base robusta e amplamente utilizada em estudos sobre humor verbal no contexto do PLN em português.

Este artigo se diferencia principalmente por seu foco exclusivo na detecção de trocadilhos em português, utilizando o corpus Puntuguese, que oferece um conjunto de dados balanceado e representativo para a tarefa. Enquanto muitos estudos anteriores tratam do humor de maneira geral, nossa pesquisa se dedica a compreender as sutilezas semânticas dos trocadilhos, um tipo de humor que exige uma interpretação mais profunda. Para isso, adotamos o modelo BERTimbau large, que foi treinado especificamente para o português, e combinamos essa abordagem com técnicas de extração de características linguísticas por expressões regulares, permitindo uma análise mais detalhada e precisa dos textos.

3. Descrição do conjunto de dados

Para a realização deste trabalho, foi utilizado o corpus Puntuguese [Inacio et al. 2024], uma coleção única de textos de trocadilhos em português, abrangendo tanto o português brasileiro quanto o europeu. Este corpus foi construído a partir de três fontes principais: o *blog* Maiores e Melhores, a página do Instagram O Sagrado Caderno das Piadas Secas, e o canal *UTC–Ultimate* Trocadilho Challenge dos Castro Brothers no YouTube. O corpus

contém 4.903 textos, classificados em duas categorias: trocadilho, com 2.450 amostras, e não trocadilho, com 2.453 amostras, o que proporciona um corpus balanceado, sem a necessidade de técnicas adicionais de balanceamento de classes. A seguir, apresentamos exemplos representativos de cada classe do corpus:

- **Exemplo de trocadilho:** "Por que a abelha foi ao médico? Porque estava com um zangão na garganta."
- **Exemplo de não trocadilho:** "Ela foi ao cinema ver um filme de drama com as amigas."

Esses exemplos ilustram a diferença entre construções textuais com jogos de palavras — que exploram ambiguidade fonológica ou semântica — e frases descritivas comuns, sem intenção humorística. Observa-se que a presença de perguntas estruturadas, como no primeiro exemplo, é frequente entre os trocadilhos, o que pode ter influenciado o modelo a superestimar o humor em determinados contextos.

As sentenças do corpus são predominantemente curtas, como evidenciado pela análise de comprimento de texto. A maioria das sentenças possui entre 50 e 100 caracteres, e a quantidade de sentenças diminui à medida que o comprimento das mesmas aumenta. Apenas um pequeno número de textos ultrapassa os 150 caracteres, indicando que os trocadilhos são, em sua maioria, textos curtos ou de tamanho médio.

A Figura 1 apresenta a distribuição dos comprimentos dos textos, em número de caracteres, separados por classe (trocadilho e não trocadilho). Observa-se que a maioria das sentenças está concentrada na faixa entre 40 e 100 caracteres, com um pico de frequência próximo aos 50 caracteres para ambas as classes. No gráfico, os textos classificados como trocadilhos são representados em azul, enquanto os não trocadilhos aparecem em verde.

A forma geral das distribuições indica uma tendência semelhante entre os dois grupos, embora os textos classificados como trocadilhos apresentem uma ligeira inclinação para comprimentos mais curtos em relação aos não trocadilhos. A frequência diminui gradualmente à medida que o comprimento do texto aumenta, refletindo o padrão esperado para conteúdos breves, como piadas ou perguntas rápidas. Essa análise sugere que o comprimento textual, embora não seja um fator decisivo isoladamente, pode contribuir como feature auxiliar em modelos de classificação automática entre trocadilhos e textos não humorísticos.

O código-fonte utilizado neste trabalho está disponível em: `https://colab.research.google.com/drive/1Vh4vqOC1m1FPCHo3oQ9Lw5jB6hKpTzkK#scrollTo=nZ_VVW4S64en`.
O corpus *Puntuguese*, utilizado como base para o treinamento do modelo, está acessível publicamente em: `https://github.com/Superar/Puntuguese`.

4. Detalhamento da Estratégia Desenvolvida

Neste artigo, a tarefa principal foi a classificação de trocadilhos utilizando modelos de Processamento de Linguagem Natural (PLN), com ênfase no modelo BERTimbau Large, uma versão treinada do BERT [Devlin et al. 2019] para a língua portuguesa.

A Figura 2 apresenta uma visão geral do pipeline desenvolvido, dividido em sete etapas: coleta do corpus, extração de features, tokenização, criação do dataset, treina-

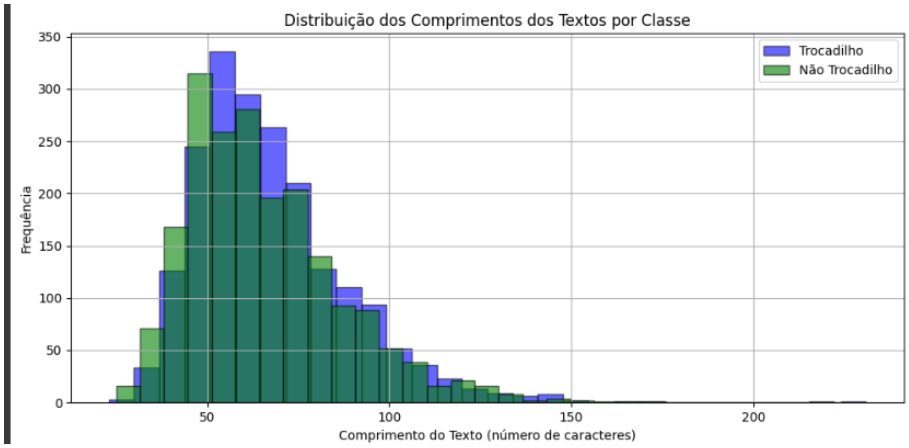

Figura 1. Histograma de comprimento de texto no Corpus Puntuguese.

mento do modelo, avaliação e visualizações. Essa estrutura modular permitiu o controle e análise sistemática de cada fase, garantindo maior transparência e reprodutibilidade.

Figura 2. Visão geral das etapas do processo de classificação de trocadilhos utilizando BERTimbau.

Uma das principais melhorias implementadas foi o ajuste no comprimento máximo das sequências de texto processadas pelo modelo. Inicialmente, o limite era de 50 tokens, o que poderia restringir a capacidade do modelo de capturar informações relevantes em sentenças mais longas. Embora as sentenças do corpus possuam, em média, cerca de 50 caracteres, o número de tokens pode ser superior, pois os *tokens* representam não apenas palavras completas, mas também pontuações e fragmentos de palavras.

A Figura 3 apresenta a distribuição do número de *tokens* por sentença no conjunto de treinamento. Observa-se que a maioria das sentenças possui entre 10 e 30 tokens, com pico de frequência entre 18 e 20 *tokens*. Esse cenário mostra que a maior parte dos dados está concentrada em sequências curtas, mas também há casos isolados com maior complexidade.

Apesar de a média estar abaixo de 50 *tokens*, optou-se por aumentar o limite para

250 *tokens*, garantindo que todas as sentenças, inclusive as mais longas e contextualmente ricas, fossem processadas sem truncamento. Essa escolha foi motivada por testes empíricos que demonstraram que o modelo se beneficia de um limite mais amplo, especialmente em trocadilhos que envolvem construções ambíguas, contextos narrativos ou repetições que ampliam o número de tokens mesmo em textos curtos.

Esse ajuste contribuiu para preservar a integridade semântica das sentenças durante o treinamento, o que se refletiu em um ganho de desempenho, permitindo ao modelo capturar melhor as nuances linguísticas e contextuais características dos trocadilhos.

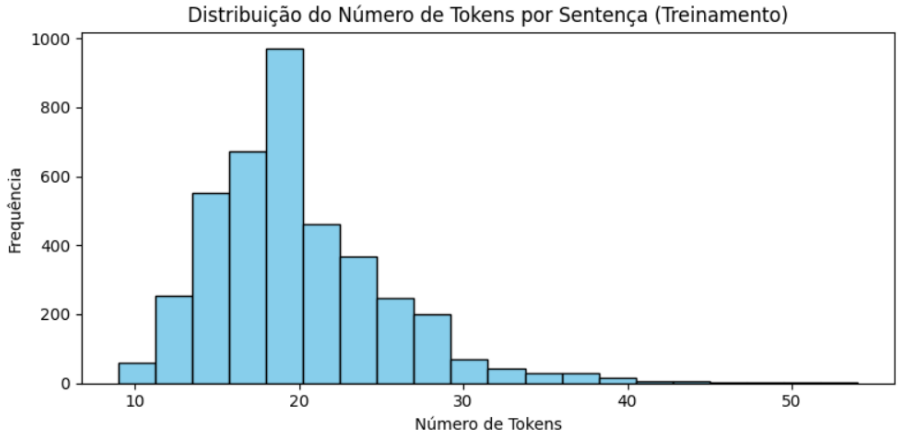

Figura 3. Distribuição do número de tokens por sentença no conjunto de treinamento.

Quanto ao balanceamento das classes, o corpus Puntuguese já era previamente balanceado, com o mesmo número de exemplos para trocadilhos e não trocadilhos. No entanto, para garantir que o modelo tratasse as classes de maneira equitativa durante o treinamento, foi utilizado o cálculo automático dos pesos de classe com base na frequência observada no conjunto de dados. Esses pesos foram aplicados diretamente na função de perda *CrossEntropyLoss* [PyTorch Documentation 2024] , tornando o modelo mais sensível a possíveis erros de classificação em ambas as categorias. A divisão dos dados foi feita da seguinte forma: treinamento: 3.990 amostras (70%), validação: 570 amostras (10%) e teste: 1.140 amostras (20%). Assim, a proporção adotada foi 70/10/20 para as etapas de treino, validação e teste, respectivamente, assegurando uma separação adequada para ajuste de parâmetros, avaliação intermediária e verificação final de desempenho.

Durante o treinamento, também foram aplicadas estratégias de regularização para evitar overfitting e melhorar a capacidade de generalização do modelo. Entre essas estratégias, destaca-se o uso de *dropout*, que ajuda a evitar o sobreajuste ajustando aleatoriamente partes da rede durante o treinamento. Também foi implementado o *early stopping*, que interrompeu o treinamento após três épocas sem melhorias significativas no desempenho, otimizando o uso de recursos computacionais. O treinamento foi realizado ao longo de 6 épocas, proporcionando tempo suficiente para o ajuste eficiente dos parâmetros.

Além disso, foi utilizado um aquecimento *(warmup)* de 100 passos no início do treinamento, o que ajudou a ajustar gradualmente a taxa de aprendizado, definida em 3e-5. Esse valor possibilitou uma aprendizagem estável, minimizando oscilações bruscas no desempenho do modelo.

As métricas utilizadas para avaliação incluíram *acurácia*, precisão, *recall* e *F1-score* ponderado. O uso do F1-score ponderado foi especialmente importante para garantir que o modelo mantivesse um bom equilíbrio entre a taxa de acertos e a sensibilidade para cada classe, mesmo diante de pequenas variações em suas distribuições.

Essas melhorias nas estratégias de pré-processamento, ajuste de hiperparâmetros, regularização e avaliação resultaram em uma solução robusta para a tarefa de classificação de trocadilhos. O uso do *BERTimbau Large*, aliado a essas técnicas avançadas, permitiu capturar as sutilezas linguísticas dos trocadilhos, oferecendo uma performance consistente em termos de *acurácia*, precisão, *recall* e *F1-score.*

Tabela 1. Configurações do Modelo de Classificação.

Componente	Valor Utilizado
Função de Perda	CrossEntropyLoss (com pesos)
Épocas de Treino	6
Batch Size	16
Learning Rate	3e-5
Early Stopping	Sim (Paciência = 3)

A Tabela 1 apresenta os principais hiperparâmetros empregados no treinamento. A arquitetura adotada incluiu 24 camadas do modelo BERTimbau Large, taxa de aprendizado de 3×10^{-5}, *batch size* de 16, função de perda com pesos balanceados (*CrossEntropyLoss*) e estratégia de parada antecipada com paciência igual a 3. Esses parâmetros foram definidos com base em testes preliminares e recomendações da literatura para modelos baseados em transformadores.

5. Análise dos resultados alcançados

O objetivo principal deste trabalho foi desenvolver um sistema de classificação capaz de identificar automaticamente trocadilhos em textos em português, utilizando o modelo *BERTimbau Large* aliado a estratégias de pré-processamento e extração de padrões linguísticos.

Ao comparar com outros trabalhos na área, o desempenho alcançado neste trabalho é satisfatório, considerando a complexidade da tarefa de detectar trocadilhos, que envolve mais do que simplesmente entender o texto, mas também a interpretação de jogos de palavras e ambiguidades semânticas. Em comparação com o estudo de Inácio et al. (2024), que apresentou um *F1-score* de 68,9% utilizando o *corpus Puntuguese*, os resultados deste trabalho são competitivos, considerando que utilizamos o mesmo conjunto de dados e uma abordagem semelhante. A tarefa de detecção de trocadilhos se mantém desafiadora por exigir compreensão contextual e identificação de nuances linguísticas.

A Tabela 2 apresenta o relatório de classificação por classe, com os principais indicadores de desempenho do modelo no conjunto de teste. Observa-se que a classe "Não Trocadilho" obteve maior recall (0,81) em comparação à classe "Trocadilho" (0,60), o que reforça a tendência do modelo em identificar com maior sensibilidade textos que não envolvem trocadilhos. Por outro lado, a classe "Trocadilho" apresentou superioridade na precisão (0,76), indicando que, embora o modelo tenha menor sensibilidade para essa

classe, tende a cometer menos erros quando a identifica. Tais observações demonstram que o modelo ainda possui margens para aprimoramento no equilíbrio entre as classes, embora o desempenho geral (acurácia = 0,70) seja considerado satisfatório.

Tabela 2. Relatório de Classificação por Classe no Conjunto de Teste.

Classe	Precisão	Recall	F1-score	Amostras
Não Trocadilho	0.67	0.81	0.73	57
Trocadilho	0.76	0.60	0.67	57
Acurácia	–	–	**0.70**	114
Média Macro	0.71	0.70	0.70	114
Média Ponderada	0.71	0.70	0.70	114

Além disso, a análise das features contextuais utilizadas pelo modelo — especialmente os padrões linguísticos relacionados a trocadilhos, extraídos por expressões regulares — indica que o modelo foi capaz de capturar nuances importantes do humor verbal. Essas features, como repetições de palavras e terminações específicas (ex.: palavras terminadas em "a" ou "o"), foram extraídas durante o pré-processamento e incorporadas como vetores adicionais no dataset. Cada exemplo, ao ser tokenizado, incluía também um vetor com contagens dessas ocorrências, fornecendo informações complementares ao modelo durante o treinamento.

Complementarmente, a Figura 4 apresenta a matriz de confusão do conjunto de teste, evidenciando o desempenho por classe. Observa-se que o modelo apresentou recall superior para a classe "Não Trocadilho" (0,81) em comparação à classe "Trocadilho" (0,60), o que indica uma tendência a classificar com maior precisão textos que não envolvem trocadilhos. Ainda assim, o equilíbrio geral das métricas aponta para um desempenho satisfatório.

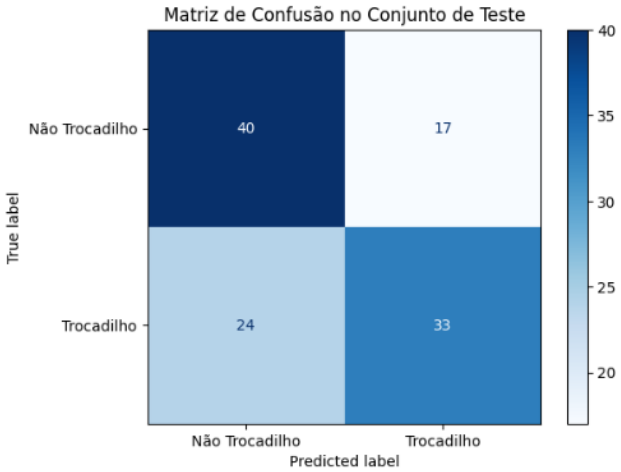

Figura 4. Matriz de Confusão no Conjunto de Teste

Além da avaliação pontual no conjunto de teste, foi realizada uma análise da curva de aprendizagem do modelo, conforme mostrado na Figura 5. A curva revela que a perda de treinamento manteve tendência de queda ao longo das épocas, enquanto a perda de

validação começou a subir a partir da quarta época — indício claro de overfitting. Essa divergência entre as curvas justifica a adoção da técnica de *early stopping*, que interrompeu o treinamento ao detectar ausência de melhoria na validação por três épocas consecutivas.

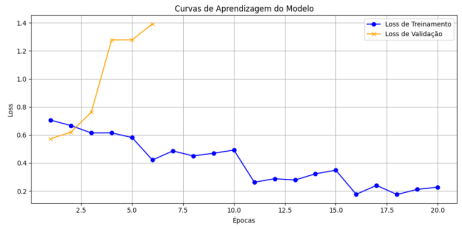

Figura 5. Curvas de Aprendizagem do Modelo (Loss de Treinamento e Validação).

Essa análise reforça a importância das estratégias de regularização empregadas e da escolha criteriosa do número de épocas de treinamento. O comportamento do modelo, tanto quantitativa quanto qualitativamente, demonstra sua capacidade de identificar trocadilhos de forma robusta, embora com espaço para melhorias em termos de generalização.

Considerações Complementares

Além das representações contextualizadas geradas pelo BERTimbau Large, o sistema desenvolvido também incorporou um conjunto de *features* linguísticas extraídas por expressões regulares, como repetições de palavras, uso de pronomes interrogativos, terminações fonológicas específicas e padrões de estrutura frasal. Embora não tenha sido realizado um experimento de ablação para isolar o impacto dessas informações, observou-se durante o desenvolvimento que sua inclusão aumentou a estabilidade das previsões, especialmente em textos curtos e ambíguos. Essas *features* funcionam como complemento ao *embedding* textual, fornecendo pistas adicionais ao classificador.

Adicionalmente, foi conduzida uma análise qualitativa dos erros cometidos pelo modelo no conjunto de teste. A seguir, são apresentados dois exemplos de classificações incorretas:

- **Texto:** "Por que o casal de patos sempre tomam banho quando discutem? Para deixar tudo em pratos limpos."
  Classe verdadeira: Não trocadilho **Classe prevista:** Trocadilho
- **Texto:** "Qual é a atriz mais dramática do mundo? A tristeza."
  Classe verdadeira: Trocadilho **Classe prevista:** Não trocadilho

Esses casos ilustram a dificuldade do modelo em lidar com jogos de palavras mais sutis, bem como sua tendência a superestimar o humor em perguntas estruturadas. O uso de construções ambíguas parece induzir o modelo a inferir incorretamente a presença de trocadilhos, mesmo quando a resposta é literal ou não contém ambiguidade semântica. Isso reforça a importância de estratégias adicionais que permitam distinguir trocadilhos implícitos de frases que apenas se assemelham ao estilo de piadas.

6. Conclusão

Neste artigo, utilizamos o modelo BERTimbau Large para a tarefa de classificação de sentenças com o objetivo de identificar trocadilhos em português. O modelo demonstrou desempenho competitivo, com bons resultados nas primeiras épocas de treinamento, reforçando seu potencial para lidar com esse tipo específico de humor verbal.

Apesar disso, observou-se a ocorrência de overfitting a partir da segunda época, caracterizada pela redução contínua da perda de treinamento e aumento progressivo da perda de validação. Esse comportamento sugere que o modelo pode ter aprendido padrões específicos dos dados de treino, sem captar plenamente a lógica subjacente aos trocadilhos. Tal limitação reforça a necessidade de estratégias complementares que favoreçam a generalização.

Entre as possibilidades de aprimoramento, destacam-se a incorporação de mais features linguísticas contextuais, como estruturas frasais e ambiguidade semântica, além da adoção de técnicas de regularização mais robustas e maior diversidade no corpus de treinamento. A utilização de embeddings semânticos mais profundos também pode contribuir para a detecção de nuances presentes nesse tipo de humor.

De modo geral, o estudo demonstrou a viabilidade de aplicar modelos de linguagem avançados à tarefa de classificação de trocadilhos, oferecendo uma base sólida para pesquisas futuras. Os resultados obtidos validam a abordagem proposta e revelam oportunidades relevantes de melhoria, especialmente no equilíbrio entre aprendizado e capacidade de generalização.

Referências

BONET, H. A.; RINCÓN, A. M.; LÓPEZ, A. M. Detection, classification and quantification of hurtful humor (huhu) on twitter using classical models, ensemble models, and transformers. In: *IberLEF@ SEPLN*. [S.l.: s.n.], 2023.

CRUZ, J. et al. In unity, there is strength: On weighted voting ensembles for hurtful humour detection. In: *IberLEF@ SEPLN*. [S.l.: s.n.], 2023.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

INACIO, M. L. et al. Puntuguese: A corpus of puns in Portuguese with micro-edits. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, 2024. p. 13332–13343. Disponível em: <https://aclanthology.org/2024.lrec-main.1167/>.

MILLER, T.; HEMPELMANN, C.; GUREVYCH, I. SemEval-2017 task 7: Detection and interpretation of English puns. In: BETHARD, S. et al. (Ed.). *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 58–68. Disponível em: <https://aclanthology.org/S17-2005/>.

PyTorch Documentation. *torch.nn.CrossEntropyLoss*. 2024. `https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html`. Acesso em: 6 abr. 2025.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian conference on intelligent systems*. [S.l.], 2020. p. 403–417.