Predição Numérica de Avaliações em Português: Comparando BERTimbau e Modelos Multilíngues

Emanuelle Marreira[1], **Tiago de Melo**[1]

Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

{erm.eng22,tmelo}@uea.edu.br

Abstract. *This study presents a comparative analysis of different BERT-based models applied to the rating prediction task. Using a dataset of over 50,000 Amazon user reviews written in Brazilian Portuguese, the performance of BERTimbau, multilingual BERT, multilingual DistilBERT, multilingual ALBERT and multilingual RoBERTa models is evaluated. BERTimbau achieved the best performance, with MAE, RMSE and AUC scores of 0.559, 0.910 and 0.933, respectively. Model effectiveness varies across product categories, with Food and Automotive showing the best results. As a contribution, this study expands the understanding of rating prediction in the context of the Portuguese Language.*

Resumo. *Este estudo analisa variantes do modelo BERT para a tarefa de predição de ratings. Com mais de 50 mil comentários de usuários da Amazon em português brasileiro, compararam-se os desempenhos dos modelos BERTimbau, BERT multilingue, DistilBERT multilingue, ALBERT multilingue e a versão multilingue do RoBERTa. O BERTimbau obteve os melhores resultados, com MAE, RMSE e AUC de 0,559, 0,910 e 0,933, respectivamente. A eficiência do modelo varia entre categorias de produtos, sendo as categorias com melhores resultados a de Alimentos e Automotivo. Como contribuição, este estudo amplia o conhecimento sobre predição de ratings no contexto da Língua Portuguesa.*

1. Introdução

Na era digital, comentários e avaliações textuais fornecem informações relevantes sobre a percepção dos consumidores em relação a produtos e serviços (de Melo et al., 2019). Entretanto, a análise dessas avaliações apresenta desafios, como o alto volume de dados, a subjetividade e a variação linguística, pois os textos não seguem um formato padronizado (Almeida Neto and de Melo, 2023). O Processamento de Linguagem Natural (PLN) é essencial para extrair informações úteis de textos de avaliação, permitindo a conversão automática de avaliações em *ratings* numéricos, o que beneficia consumidores, empresas e sistemas de recomendação (Li et al., 2022b).

O modelo *Transformer* (Vaswani et al., 2017) se tornou a base para os avanços em PLN, destacando-se pelo desempenho expressivo em tarefas como classificação de textos (Gardazi et al., 2025). Embora o português seja uma das cinco línguas mais utilizadas na internet (Pereira, 2021), a aplicação dessa arquitetura em tarefas como predição de *ratings* (*rating prediction*) em comentários nesse idioma ainda é pouco explorada. Assim, este estudo avalia diferentes versões do modelo *Bidirectional Encoder Representations from Transformers* (BERT) no problema de *rating prediction* de comentários em

português brasileiro de usuários da Amazon[1]. A plataforma foi escolhida pela popularidade e diversidade de categorias de produto, o que enriquece a variedade semântica dos comentários analisados. Além disso, como contribuição, o código desenvolvido neste trabalho e o conjunto de dados utilizados são disponibilizados no GitHub[2].

A Figura 1 ilustra um exemplo de comentário real do site, onde um modelo deve prever a nota atribuída ao produto a partir exclusivamente do texto da avaliação. Neste caso, o modelo deveria apresentar o resultado de 2 estrelas. Esse processo enfrenta desafios devido à alta subjetividade presente nas opiniões dos usuários, como o caso apresentado, onde o usuário escreve elogios e críticas no mesmo texto.

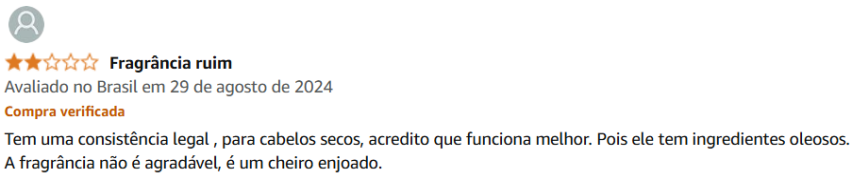

Figura 1. Comentário real de um usuário da Amazon.

Para aprofundar a investigação, este estudo propõe duas perguntas de pesquisa:

PP1: Alguma versão do BERT se destaca na tarefa de *rating prediction* entre diferentes categorias de produtos?

PP2: A melhor configuração do modelo mantém seu desempenho ao considerar separadamente essas categorias?

Em resposta à PP1, o modelo BERTimbau apresentou desempenho superior em todas as métricas, destacando-se entre as diferentes variações de BERT. A partir disso, conduziu-se um estudo para responder à PP2, em que o modelo BERTimbau foi aplicado a conjuntos de dados de diferentes categorias de produto. Nesse estudo, identificou-se que as categorias de Alimentos e Automotivo alcançaram os melhores resultados. Para aprofundar a análise, realizou-se um estudo com métricas textuais dos comentários dessas categorias, o qual indicou que elas apresentam, em média, menor número de palavras e sentenças por comentário, o que deve favorecer a tarefa de classificação de texto.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3, a metodologia; a Seção 4, os experimentos e resultados; e a Seção 5, as conclusões e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

2.1. *Rating Prediction*

Chambua and Niu (2021) apresentam uma revisão sobre abordagens que usam análise de textos de comentários para *rating prediction*, incluindo técnicas de análise de sentimentos e extração de tópicos. Trabalhos em sistemas de recomendação abordam o problema de *rating prediction* geralmente combinando informações do usuário, do produto e do histórico de interações para obter previsões mais precisas (Shi et al. (2020), Li et al.

[1]https://www.amazon.com.br/

[2]https://github.com/emanuellemarreira/rating-prediction-with-bert

(2022a)). Por outro lado, há abordagens que se baseiam apenas no conteúdo dos comentários, o que torna a tarefa mais desafiadora, pois dependem das nuances do texto para predizer *ratings*, sem informações contextuais adicionais.

Nesse contexto, Hanić et al. (2024) tratam *rating prediction* como um problema de classificação multiclasse, investigando a relação entre resenhas e avaliações numéricas de restaurantes veganos e vegetarianos do TripAdvisor, comparando representações de texto *Term Frequency-Inverse Document Frequency* (TF-IDF) e GloVe. Na perspectiva de regressão, tendo em vista que as classes de estrelas variam de 1 a 5, Khan et al. (2022) propõem uma abordagem que divide os dados em subconjuntos positivos e negativos, aplicando modelos como o XGBoost. Ainda na perspectiva de regressão, Antonio et al. (2018) prevê as *ratings* de avaliações online de hotéis, usando *n-grams* e TF-IDF, a partir de fontes como Booking e TripAdvisor, abrangendo múltiplos idiomas. Com foco em múltiplos idiomas, Hossain (2021) propõe uma metodologia para prever classificações de avaliações de produtos em Bangla, utilizando algoritmos de aprendizado de máquina e TF-IDF para extração de características.

Apesar do aumento de modelos como o BERT, a maioria dos estudos concentra-se em textos em inglês ou espanhol, geralmente utilizando metadados. Não foram encontrados trabalhos que abordem a tarefa de *rating prediction* em português brasileiro exclusivamente com base nos textos, usando variantes do BERT.

2.2. BERT

Estudos recentes exploram as oportunidades e desafios dos modelos baseados em *Transformer* na área de PLN. Pak et al. (2024) apresentam uma análise sistemática das diversas técnicas de representação vetorial de palavras, destacando o BERT como um marco na evolução dos *embeddings*, enfatizando que sua arquitetura supera métodos tradicionais com representações fixas e unidirecionais. Gardazi et al. (2025) revisam as aplicações do BERT em várias tarefas de PLN, abordando desafios e perspectivas dos modelos. Além disso, Cunha et al. (2025) comparam BERT e outros modelos modernos com técnicas clássicas de classificação, considerando desempenho e custo computacional.

Assim, para o problema de *rating prediction*, na análise de sentimentos, diversas técnicas podem ser exploradas. Apesar dos avanços em PLN, a pesquisa concentra-se em dados em inglês, evidenciando a falta de estudos e dados em português, especialmente brasileiro. Este trabalho busca preencher essa lacuna, examinando o desempenho de diferentes versões de modelos BERT na classificação multiclasse de *rating prediction*. Até onde se sabe, este é o primeiro estudo que aborda o problema dessa forma no contexto da língua portuguesa.

3. Materiais e Métodos

3.1. Conjunto de Dados

O conjunto de dados da pesquisa contém comentários de produtos de diversas categorias do site de comércio eletrônico da Amazon Brasil, referentes ao período de 2021 a 2024, obtidos por meio de coleta manual e de técnicas de *web scraping*. Apenas comentários em português foram considerados. A pesquisa utilizou somente o texto do comentário e seu respectivo *rating*, um número inteiro de 1 a 5, atribuído pelo usuário.

A Tabela 1 resume as estatísticas do conjunto de dados. Inicialmente, o conjunto de dados apresentava um desbalanceamento entre as classes de *rating*, motivo pelo qual foi aplicado um *undersampling* com o objetivo de igualar a quantidade de comentários por classe em cada categoria de produto. Observa-se maior volume de comentários nas categorias *Livros* e *Moda*, enquanto *Computadores* e *Pets* são menos representadas. Essa variação reflete a popularidade das categorias, com destaque para *Livros*, impulsionada pelo hábito dos leitores de avaliarem suas leituras na Amazon.

Tabela 1. Distribuição de comentários e notas de avaliação por categoria.

Categoria	**Comentários por Avaliações**					**Total**
	1	**2**	**3**	**4**	**5**	**de Comentários**
Automotivo	873	873	873	873	873	4.365
Bebê	1.057	1.057	1.057	1.057	1.057	5.285
Celulares	867	867	867	867	867	4.335
Alimentos	742	742	742	742	742	3.710
Jogos	1.217	1.217	1.217	1.217	1.217	6.085
Computadores	185	185	185	185	185	925
Livros	2.259	2.259	2.259	2.259	2.259	11.295
Moda	1.443	1.443	1.443	1.443	1.443	7.215
Pets	445	445	445	445	445	2.225
Brinquedos	1.205	1.205	1.205	1.205	1.205	6.025
TOTAL	**10.293**	**10.293**	**10.293**	**10.293**	**10.293**	**51.465**

3.2. Modelos Utilizados

Para este estudo, utilizou-se uma variedade de modelos de linguagem pré-treinados disponíveis na plataforma Hugging Face[3]. Os modelos selecionados foram testados nas versões *base* e *cased*, com distinção entre maiúsculas e minúsculas, garantindo uma comparação justa, pois nem todos tinham versões *uncased*. Foram considerados os modelos: BERTimbau, BERT multilingue, DistilBERT multilingue, ALBERT multilingue e RoBERTa multilingue. Esses modelos foram escolhidos por sua relevância na análise do impacto das variantes do BERT no desempenho em tarefas em língua portuguesa.

3.3. Métricas de Avaliação

Como o problema de *rating prediction* pode ser visto como uma tarefa de regressão, foram usadas as métricas RMSE e MAE, comumente empregadas nessa tarefa, para avaliar os resultados. Assim, é possível analisar a tendência de erros dos modelos entre as cinco classes de *rating*, levando em conta a subjetividade que pode gerar confusão entre classes semelhantes na classificação. Seja N o número total de observações, y_i os valores reais e $\hat{y}_i$ os valores preditos. As métricas RMSE e MAE são definidas como:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (1) \qquad MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (2)$$

A MAE (2) mede o erro absoluto médio nas unidades originais dos dados, enquanto a RMSE (1) enfatiza erros maiores ao elevar as diferenças ao quadrado, tornando-

[3]https://huggingface.co/

se mais sensível a *outliers*. Em ambos os casos, valores menores indicam melhor desempenho do modelo.

Adicionalmente, foi calculada a métrica *Area Under the Curve* (AUC), que representa a área sob a curva ROC e indica a capacidade do modelo de separar classes, variando de 0 a 1, onde valores próximos a 1 sugerem melhor desempenho. Contudo, essa métrica é geralmente aplicada em problemas de classificação binária. Assim, seguindo abordagens similares às de Kang et al. (2023) e de Melo (2022), as classes 4 e 5 foram agrupadas como positivas, e as demais como negativas, transformando o problema em uma tarefa binária. Essa estratégia permite avaliar de forma mais direta como o modelo distingue avaliações altamente positivas das demais, uma prática comum em estudos que buscam reduzir a ambiguidade de classes próximas (Antonio et al., 2018).

4. Experimentos

4.1. Comparação de versões de BERT para *rating prediction*

Para responder à primeira pergunta de pesquisa (PP1), foram realizados experimentos com os modelos da Seção 3.2 para identificar a melhor versão de BERT para a tarefa de *rating prediction* em português. O *fine-tuning* dos modelos usou a classe *Trainer*, da biblioteca Transformers da Hugging Face. Os dados foram estratificados por categoria de produto e por classe de *rating*, sendo divididos em 80% para treinamento e 20% para teste, utilizando uma estratégia de *hold-out*, adequada devido à quantidade suficiente de dados disponíveis. A otimização de hiperparâmetros foi feita com a biblioteca Optuna[4], variando parâmetros como *learning rate*, *batch size*, número de *epochs* e *weight decay* para minimizar a métrica RMSE. A Tabela 2 apresenta os resultados.

Tabela 2. Resultados dos modelos com comentários de todas as categorias.

Modelo	MAE	RMSE	AUC
BERTimbau	**0,559**	**0,910**	**0,933**
BERT multilingue	0,616	0,962	0,906
DistilBERT multilingue	0,619	0,980	0,921
ALBERT multilingue	0,711	1,096	0,893
RoBERTa multilingue	0,562	**0,910**	0,930

O modelo BERTimbau obteve os melhores resultados em todas as métricas, indicando que esta é a melhor versão do BERT para a tarefa de *rating prediction* em português. Esse resultado era esperado, pois o BERTimbau é pré-treinado especificamente para o português brasileiro, diferentemente das versões multilingues.

4.2. Comparação de diferentes categorias de produto para *rating prediction*

A segunda pergunta de pesquisa (PP2) investiga se a melhor configuração do modelo mantém a eficiência ao ser treinada em conjuntos de dados por categoria de produto. Para isso, o modelo BERTimbau foi treinado visando minimizar a métrica RMSE, com a divisão de treino e teste, ambos por categoria, totalizando 10 subconjuntos de dados. A Tabela 3 apresenta o desempenho de cada categoria. A categoria *Automotivo* teve o melhor resultado na métrica AUC, enquanto *Alimentos* apresentou o melhor desempenho nas métricas MAE e RMSE.

[4]https://optuna.org/

Tabela 3. Resultados (PP2).

Categoria	MAE	RMSE	AUC
Automotivo	0,547	0,874	**0,943**
Bebê	0,526	0,850	0,926
Livros	0,588	0,940	0,923
Celulares	0,703	1,051	0,905
Moda	0,597	0,914	0,931
Alimentos	**0,509**	**0,834**	0,927
Jogos	0,577	0,888	0,919
Computadores	0,708	1,058	0,929
Pets	0,566	0,887	0,917
Brinquedos	0,591	0,927	0,936

Tabela 4. Métricas linguísticas.

Categoria	MNPC	MNSC	FRE	RTT
Automotivo	18,090	1,795	51,153	0,938
Bebê	17,834	1,821	53,942	0,942
Celulares	24,026	2,019	47,594	0,921
Alimentos	17,166	1,830	55,896	0,941
Jogos	22,955	1,878	51,224	0,923
Computadores	29,466	2,464	48,698	0,914
Livros	27,429	2,204	52,596	0,916
Moda	16,028	1,772	50,465	0,946
Pets	20,271	1,904	55,107	0,935
Brinquedos	19,427	1,922	52,209	0,938

Para entender essa dificuldade, foi realizada uma análise estatística dos comentários nas diferentes categorias. Foram calculadas as métricas: Média do Número de Palavras por Comentário (MNPC), Média do Número de Sentenças por Comentário (MNSC), média do Índice de Legibilidade de Flesch (FRE) dos comentários da categoria, que avalia a facilidade de leitura do texto (Eleyan et al., 2020), e a Razão *Type-Token* (RTT), que mede a diversidade lexical (Kettunen, 2014). A Tabela 4 apresenta os resultados para cada categoria analisada.

Embora essas métricas ajudem a analisar a complexidade dos textos, não foi observada uma correlação clara entre esses indicadores e o desempenho dos modelos na tarefa de predição de *ratings*. As piores performances nas métricas MAE e RMSE ocorreram nas categorias de *Computadores* e *Celulares*, que têm as maiores médias de palavras (MNWC) e sentenças por comentário (MNSC). Em contrapartida, *Alimentos* e *Bebê*, com valores médios mais baixos, obtiveram melhores resultados. Na métrica AUC, a categoria *Celulares* apresenta novamente um desempenho inferior, enquanto *Automotivo*, com bons resultados, também está entre as categorias com menores valores de MNWC e MNSC. Além disso, todas as categorias exibem RTT acima de 0,90, indicando vocabulário variado, o que pode dificultar a tarefa de classificação. Outro fator que pode contribuir para o desempenho inferior dos modelos é o FRE, pois os baixos valores sugerem textos entre "um pouco difíceis" e "difíceis" de ler.

5. Conclusões

Este estudo comparou versões do modelo BERT para o problema de *rating prediction* com comentários de usuários da Amazon em português brasileiro. Os resultados indicaram que o BERTimbau pré-treinado teve o melhor desempenho em todas as métricas, utilizando um conjunto de dados com comentários de várias categorias. Na análise por categoria, o BERTimbau apresentou os melhores resultados em *Alimentos* e *Automotivo*. Uma limitação do estudo é o foco exclusivo na Amazon Brasil, o que pode restringir a generalização para outros domínios ou idiomas. Para trabalhos futuros, sugere-se incluir comentários em diferentes idiomas e utilizar modelos mais recentes de PLN para avaliar o problema de *rating prediction* sob nova perspectiva.

Referências

Almeida Neto, J. and de Melo, T. (2023). Exploring supervised learning models for multi-label text classification in brazilian restaurant reviews. *Anais do Encontro Nacional de*

Inteligência Artificial e Computacional (ENIAC), pages 126–140.

Antonio, N., de Almeida, A. M., Nunes, L., Batista, F., and Ribeiro, R. (2018). Hotel online reviews: creating a multi-source aggregated index. *International Journal of Contemporary Hospitality Management*, 30(12):3574–3591.

Chambua, J. and Niu, Z. (2021). Review text based rating prediction approaches: preference knowledge learning, representation and utilization. *Artificial Intelligence Review*, 54:1171–1200.

Cunha, W., Rocha, L., and Gonçalves, M. A. (2025). A thorough benchmark of automatic text classification: From traditional approaches to large language models. *arXiv preprint arXiv:2504.01930*.

de Melo, T. (2022). Sentilexbr: An automatic methodology of building sentiment lexicons for the portuguese language. *Journal of Information and Data Management*, 13(3).

de Melo, T., da Silva, A. S., de Moura, E. S., and Calado, P. (2019). Opinionlink: Leveraging user opinions for product catalog enrichment. *Information Processing & Management*, 56(3):823–843.

Eleyan, D., Othman, A., and Eleyan, A. (2020). Enhancing software comments readability using flesch reading ease score. *Information*, 11(9):430.

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., and Alshemaimri, B. (2025). Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):1–49.

Hanić, S., Bagić Babac, M., Gledec, G., and Horvat, M. (2024). Comparing machine learning models for sentiment analysis and rating prediction of vegan and vegetarian restaurant reviews. *Computers*, 13(10):248.

Hossain, M. I. e. a. (2021). Rating prediction of product reviews in bangla using machine learning. In *Proc. Int. Conf. on AI and Mechatronics Systems (AIMS)*, pages 1–6. IEEE.

Kang, W.-C., Ni, J., Mehta, N., Sathiamoorthy, M., Hong, L., Chi, E., and Cheng, D. Z. (2023). Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Khan, R. A., Mannan, A., and Aslam, N. (2022). Prediction of product rating based on polarized reviews using supervised machine learning. *VFAST Transactions on Software Engineering*, 10(4):01–09.

Li, J., Wang, Y., and Tao, Z. (2022a). A rating prediction recommendation model combined with the optimizing allocation for information granularity of attributes. *Information*, 13(1):21.

Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., and Yu, Z. (2022b). Text mining of user-generated content (ugc) for business applications in e-commerce: A systematic review. *Mathematics*, 10(19):3554.

Pak, A., Ziyaden, A., Saparov, T., Akhmetov, I., and Gelbukh, A. (2024). Word embeddings: A comprehensive survey. *Computación y Sistemas*, 28(4):2005–2029.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Shi, W., Wang, L., and Qin, J. (2020). Extracting user influence from ratings and trust for rating prediction in recommendations. *Scientific reports*, 10(1):13592.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.