Rating Prediction in Brazilian Portuguese Reviews: An Approach Based on Textual Features

Emanuelle Marreira[1], **Miguel de Oliveira**[2], **Tiago de Melo**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Avenida Darcy Vargas, nº 1200 – Parque 10 de Novembro – Manaus – AM – Brasil

[2]Instituto de Computação – Universidade Federal do Amazonas
Av. Rodrigo Otávio, nº 6200 – Coroado I – Manaus – AM – Brasil

erm.eng22@uea.edu.br, miguel.oliveira@icomp.ufam.edu.br, tmelo@uea.edu.br

Abstract. *This paper investigates rating prediction in user reviews from Amazon written in Brazilian Portuguese, leveraging textual features and machine learning models. We propose and analyze different groups of textual features, with experimental results highlighting the crucial role of lexical information in this task. Model performance varies across product categories, with higher accuracy observed in domains with a more homogeneous vocabulary. As a contribution, this study reinforces the significance of textual representations in automated review analysis and advances the understanding of rating prediction within the context of the Portuguese language.*

Resumo. *Este trabalho explora a predição de ratings em avaliações de usuários da Amazon escritas em português brasileiro, utilizando features textuais e modelos de aprendizado de máquina. Diferentes grupos de características textuais foram propostos e analisados, e os experimentos indicam que informações léxicas desempenham um papel fundamental na tarefa. A eficácia dos modelos varia entre categorias de produtos, sendo maior em domínios com vocabulário mais homogêneo. Como contribuição, este estudo reforça a importância de representações textuais na análise automatizada de avaliações e amplia o conhecimento sobre predição de ratings no contexto da língua portuguesa.*

1. Introdução

Na era digital, plataformas *online*, como redes sociais e sites de *e-commerce*, geram uma enorme quantidade de dados das interações dos usuários. Comentários e avaliações textuais refletem as percepções dos consumidores sobre produtos e serviços [de Melo et al. 2019]. Entretanto, a análise dessas avaliações enfrenta desafios, como o alto volume de dados, a subjetividade e a variação linguística, uma vez que os textos não seguem um formato padronizado [de Almeida Neto and de Melo 2023].

O Processamento de Linguagem Natural (PLN) é um campo de pesquisa que investiga métodos para extrair informações úteis de textos, transformando grandes volumes de conteúdo não estruturado em conhecimento acionável. No contexto de avaliações de produtos e serviços, a pesquisa sobre métodos automáticos que traduzem avaliações textuais em *ratings* numéricos não apenas contribui para o avanço científico, mas também oferece aplicações práticas relevantes. Para consumidores, esses métodos facilitam a

interpretação de *feedbacks* e a comparação entre alternativas, auxiliando decisões de compra mais informadas. Para empresas, permitem monitorar em larga escala a percepção do público, identificar rapidamente problemas recorrentes e compreender tendências de satisfação ou insatisfação ao longo do tempo [Li et al. 2022b].

Estudos indicam que diferentes características textuais (*features*) influenciam a classificação de texto, incluindo aspectos sintáticos e gramaticais [Chenlo and Losada 2014, Anchiêta et al. 2021, de Oliveira and de Melo 2021]. Embora modelos Transformer sejam o estado da arte em PLN, exigem alta capacidade de memória [Stankevičius and Lukoševičius 2024]. Assim, o melhor desempenho ocorre quando o conjunto de *features* é pequeno, mas informativo [A. Semary et al. 2024]. Diante disso, torna-se relevante investigar quais *features* textuais são mais eficazes para *rating prediction*, especialmente em português, idioma amplamente usado na internet, mas pouco explorado nesse contexto [Pereira 2021].

Neste estudo, avalia-se um conjunto de 58 *features* textuais aplicadas a diversos algoritmos de aprendizagem de máquina para o problema de *rating prediction* de comentários de usuários da Amazon[1], escritos em português brasileiro. A Figura 1 ilustra um exemplo real do site, onde um modelo deve prever a nota do produto com base apenas no texto da avaliação. Neste caso, o modelo deveria atribuir 2 estrelas. Esse processo é desafiador devido à elevada subjetividade nas opiniões dos usuários, como no caso em que elogios e críticas aparecem no mesmo texto.

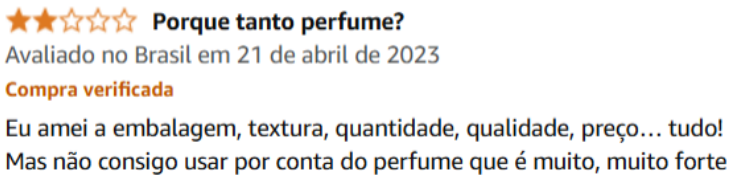

Figura 1. Comentário real de um usuário da Amazon.

Para aprofundar essa investigação, buscam-se respostas para as seguintes perguntas de pesquisa:

PP1: Qual algoritmo de aprendizagem de máquina apresenta melhor desempenho em *rating prediction* usando apenas *features* textuais, por categoria?

PP2: Existe um grupo de *features* que seja mais relevante para essa tarefa?

PP3: Como a combinação das *features* impacta o desempenho do classificador?

PP4: A melhor configuração do modelo encontrada mantém seu desempenho ao comparar comentários de diferentes categorias de produtos da Amazon?

Para responder à PP1, foram avaliados diversos modelos supervisionados em diferentes categorias de produtos, identificando-se o *eXtreme Gradient Boosting* (XGB) como o mais eficaz. Para PP2, investigaram-se conjuntos individuais de *features* textuais e realizou-se um estudo de ablação (*ablation study*), evidenciando-se que *features* de léxico são as mais relevantes. A PP3 foi analisada por meio do algoritmo *Recursive Feature Elimination* (RFE), uma técnica de seleção de variáveis que remove iterativamente

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

as menos relevantes, até encontrar o conjunto que otimiza o desempenho do modelo. Por fim, em relação à PP4, avaliou-se o desempenho do modelo em diferentes categorias, observando-se que a categoria de produto *Moda* apresentou os melhores resultados, enquanto a categoria de *Livro* se mostrou a mais desafiadora. O código e os dados utilizados estão disponíveis em um repositório[2].

2. Trabalhos Relacionados

2.1. *Rating Prediction*

[Chambua and Niu 2021] apresentam uma revisão abrangente sobre abordagens de análise de textos para a tarefa de *rating prediction*. Os autores destacam o uso de conhecimento variado, como tópicos, análise de sentimentos, aspectos linguísticos e informações semânticas, que ajudam a lidar com problemas como escassez de dados e *cold-start*.

Sistemas de recomendação lidam com *rating prediction* combinando informações do usuário, produto e do histórico de interações para realizar previsões mais precisas [Li et al. 2022a]. Outras abordagens baseiam-se apenas no conteúdo dos comentários, tornando a tarefa desafiadora, pois dependem das nuances do texto para prever *ratings*, sem informações contextuais adicionais. Nesse contexto, abordagens baseadas apenas no conteúdo textual tratam *rating prediction* como classificação multiclasse e usam representações como n-gramas e LSI [Chambua and Niu 2021]. Para a extração de características, utilizam-se n-gramas e *Latent Semantic Indexing* (LSI), uma técnica que identifica tópicos nos comentários. Na mesma abordagem dos comentários, [Hanić et al. 2024] investigam a relação entre resenhas escritas e avaliações numéricas de restaurantes veganos e vegetarianos no TripAdvisor, comparando representações de texto *Term Frequency-Inverse Document Frequency* (TF-IDF) e GloVe.

Além de classificação multiclasse, o problema de *rating prediction* pode ser tratado como uma tarefa de regressão, já que as classes de estrelas são valores numéricos de 1 a 5. [Khan et al. 2022] abordam *rating prediction* como um problema de regressão, propondo uma abordagem baseada em avaliações polarizadas, dividindo o conjunto em subconjuntos positivos e negativos. Após essa divisão, são aplicados algoritmos de aprendizagem de máquina e aprendizado profundo, sendo o XGB o modelo mais consistente, com os menores valores de *Root Mean Squared Error* (RMSE) em todos os conjuntos. [Antonio et al. 2018] também propuseram um modelo de regressão para prever os *ratings* de avaliações online de hotéis a partir de múltiplas fontes. O estudo utiliza dados do Booking e Tripadvisor, abrangendo textos em português de Portugal, espanhol e inglês, e normaliza as diferentes escalas de avaliação. Os autores exploram representações textuais como TF-IDF e n-gramas.

Na análise de *rating prediction* em diferentes idiomas, [Hossain et al. 2021] propõem uma metodologia para prever classificações de avaliações de produtos no idioma Bangla (Língua Bengali), utilizando algoritmos de aprendizagem de máquina e TF-IDF com n-gramas para extração de características.

Não foram identificados estudos que abordem o problema de *rating prediction* em português brasileiro usando apenas textos de comentários como entrada para modelos de

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

aprendizagem de máquina. A maioria das pesquisas foca em inglês e espanhol, frequentemente combinando informações contextuais, como metadados do usuário e histórico de interações, para melhorar a previsão de *ratings*. Este trabalho, em contraste, avalia quais aspectos textuais são mais relevantes para a predição de *ratings* em português, explorando um conjunto diversificado de *features* linguísticas, estruturais e semânticas aplicadas a modelos tradicionais de aprendizagem de máquina. A pesquisa contribui para o avanço da área na interseção entre PLN e análise de sentimentos em avaliações de produtos.

2.2. *Features* textuais

No contexto da língua portuguesa, [Pereira 2021] apresenta uma revisão abrangente das pesquisas que abordam as diferentes tarefas da análise de sentimentos na língua portuguesa, bem como as ferramentas utilizadas. O autor destaca o uso de técnicas de aprendizagem de máquina que extraem *features* textuais, como presença e frequência de termos, POS, palavras de opinião e negações para classificar sentimentos. O trabalho de [Anchiêta et al. 2021] aborda a detecção de ironia em textos em português brasileiro. Extraiu-se do texto o que chamam de *features* superficiais, como o número de entidades nomeadas, presença de símbolos, número de *emojis* e palavras frequentes, além de *embeddings*. Os modelos incluíram SVM e MLP (*Multi-Layer Perceptron*). Os resultados mostraram que a abordagem de *features* superficiais com SVM teve a melhor performance em tuítes, alcançando a primeira posição, enquanto a abordagem de *embeddings* com MLP ficou em segundo lugar, sugerindo que as *features* superficiais discriminam melhor textos irônicos de não irônicos que os *embeddings*.

O trabalho de [de Oliveira and de Melo 2021] identifica sentenças subjetivas em português usando diferentes *features* textuais e algoritmos de aprendizagem de máquina. A metodologia inclui a extração de *features* de POS, *features* léxicas e *features* sintáticas, aplicando os algoritmos *Gradient Boosting Trees*, *Logistic Regression*, *Random Forest*, SVM e *AutoGluon*. Os resultados indicaram que as tags POS são o grupo de *features* mais relevante para a classificação de subjetividade, e que o número de adjetivos foi a *feature* mais importante em três dos quatro conjuntos de dados. Os autores concluíram que a seleção de *features* melhora a precisão da classificação.

Embora abordagens com *embeddings* contextuais e modelos de linguagem de larga escala (LLMs) frequentemente apresentem ganhos em tarefas de classificação de texto, seus custos de implantação e operação são substancialmente maiores [Stankevičius and Lukoševičius 2024]. Neste trabalho, delimitou-se o escopo à avaliação de características textuais em modelos tradicionais, estabelecendo um *baseline* interpretável e reprodutível para o português brasileiro.

Portanto, para o problema de *rating prediction*, relacionado à análise de sentimentos, existem várias técnicas a serem exploradas. Apesar dos avanços em PLN, a literatura foca majoritariamente em dados em inglês, revelando a escassez de estudos sobre o português, especialmente o brasileiro. Assim, este trabalho visa preencher essa lacuna ao investigar o impacto de um conjunto de *features* textuais na tarefa de *rating prediction*, utilizando modelos tradicionais de aprendizagem de máquina. Até onde se tem conhecimento, este é o primeiro trabalho a abordar o problema dessa forma na língua portuguesa.

3. Metodologia

A Figura 2 apresenta um fluxograma que ilustra os passos da pesquisa. A primeira fase consiste na coleta de dados, seguida pela preparação destes. Em seguida, as *features* textuais são extraídas. Finalmente, os dados são utilizados para treinar modelos de Aprendizagem de Máquina em tarefas de classificação.

Figura 2. Fluxograma de passos da pesquisa.

3.1. Coleta e preparação de dados

O conjunto de dados consiste em comentários de produtos de várias categorias da Amazon Brasil, de 2021 a 2024, coletados manualmente e com técnicas de *web scraping*. Consideram-se apenas comentários em português brasileiro, usando o texto do comentário e seu *rating* correspondente, um número de 1 a 5 dado pelo usuário.

A Tabela 1 resume os dados da pesquisa. Observa-se desbalanceamento entre as categorias. *Livros* e *Moda* têm mais dados, enquanto *Computadores* e *Pets* são menos representadas. Algumas categorias são menos populares, resultando em menos comentários, enquanto *Livros* recebe muitas avaliações devido ao hábito dos leitores da Amazon. Para uma comparação justa entre algoritmos, aplicou-se a técnica de *undersampling* aleatória, reduzindo amostras das classes majoritárias (4 e 5) para igualar às classes minoritárias (1, 2 e 3) e equilibrar a proporção, mitigando o viés nos modelos.

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

Categoria	**Comentários por Avaliações**					**Total de Comentários**
	1	**2**	**3**	**4**	**5**	
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. Extração de *features* textuais

Para a extração de *features* textuais, foi produzido um protocolo de extração com *features* fundamentado nos trabalhos de [Chenlo and Losada 2014, de Oliveira and de Melo 2021] e outras *features* propostas pelos próprios autores. Foi realizada uma *pipeline* de extração

de *features* de texto a partir de recursos disponíveis em bibliotecas como spaCy[3], Scikit-Learn[4], SenticNet[5] e Hunspell[6]. A Tabela 2 apresenta um resumo das 58 *features* que foram utilizadas nos experimentos, onde as *features* estão organizadas em 7 grupos.

Tabela 2. Descrição dos grupos de *features* de texto.

Grupo	Descrição
Part-of-speech (POS)	Conta a ocorrência de cada categoria gramatical identificada pelo spaCy. As categorias incluem: adjetivo (F1), advérbio (F2), artigo (F3), conjunção coordenativa (F4), conjunção subordinativa (F5), interjeição (F6), marcador de pontuação (F7), número cardinal (F8), partículas (F9), preposições (F10), pronome (F11), pronome próprio (F12), símbolo (F13), substantivo (F14), superlativo (F15), verbo (F16), verbo auxiliar (F17), comparativo (F18) e tokens não categorizáveis (F19).
Padrões sintáticos (SYNT)	Conta a ocorrência dos padrões sintáticos [Chenlo and Losada 2014]: *Adjetivo* $\rightarrow$ *Substantivo* (F20), *Advérbio* $\rightarrow$ *Adjetivo* $\rightarrow$ *Não-Substantivo* (F21), *Adjetivo* $\rightarrow$ *Adjetivo* $\rightarrow$ *Não-Substantivo* (F22), *Substantivo* $\rightarrow$ *Adjetivo* $\rightarrow$ *Não-Substantivo* (F23), *Advérbio* $\rightarrow$ *Verbo* (F24) seguido por um dos tempos verbais: particípio, gerúndio, passado simples ou imperfeito.
Léxico (LEX)	Conta a ocorrência de: Termos subjetivos (F25), palavras positivas (F26), palavras negativas (F27), marcadores de pontuação (!, ?, ...) (F28), quantidade de frases positivas (F29), quantidade de frases negativas (F30); Mede a proporção em relação ao texto total de: palavras positivas (F31), palavras negativas (F32), marcadores de pontuação (F33) e termos subjetivos no texto (F34). O léxico utilizado é o SentiProdBR [de Melo 2021].
Conceito (CONC)	Inclui as seguintes métricas extraídas do SenticNet: Soma de *pleasantness* (F35), média de *pleasantness* (F36), soma de *sensitivity* (F37), média de *sensitivity* (F38), soma de *aptitude* (F39), média de *aptitude* (F40), soma de *attention* (F41), média de *attention* (F42), soma da polaridade geral do comentário (F43) e média da polaridade geral do comentário (F44).
Estruturais (EST)	Mede características estruturais do texto, incluindo: Número total de caracteres (F45), número de palavras (F46), número de frases (F47), número de palavras iniciadas por letra maiúscula (F48), número total de letras maiúsculas (F49), proporção de palavras iniciadas por maiúscula em relação ao total de palavras (F50) e proporção de letras maiúsculas em relação ao total de caracteres (F51).
Twitter/X (TWT)	Conta a ocorrência de termos típicos de redes sociais, incluindo: Número de palavras alongadas (ex: "muuuito") (F52), número de gírias específicas do Twitter/X (F53), número de palavras de negação (ex: "não", "nunca") (F54), número de emojis (F55) e número de emoticons (F56). As *features* de emoji e emoticon foram extraídas com a pontuação de sentimento proposta por [Kimura and Katsurai 2018] e [Hogenboom et al. 2015], respectivamente.
Miscelânea (MCL)	Conta a ocorrência do número de Entidades Nomeadas (ex: nomes próprios, marcas, locais) (F57) e quantidade de palavras escritas corretamente (F58), conforme as bibliotecas spaCy e o dicionário da biblioteca Hunspell.

[3]https://spacy.io/
[4]https://scikit-learn.org/stable/
[5]https://sentic.net/
[6]https://hunspell.github.io/

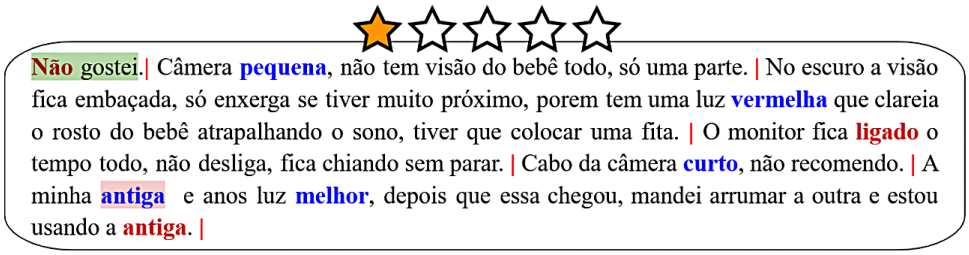

Figura 3. Exemplo de comentário real de um usuário da Amazon.

Para ilustrar a extração de *features* textuais, a Figura 3 apresenta um comentário real de um usuário da Amazon. A figura mostra exemplos de algumas *features* devido à limitação do número de páginas. A *feature* F24 aparece no texto grifado em verde, contendo um advérbio e um verbo. No mesmo trecho, a palavra de negação "Não", destacada em vermelho, é identificada pela *feature* F54. As palavras em azul representam adjetivos identificados pela *feature* F1, totalizando cinco ocorrências. Os três termos em vermelho foram reconhecidos como negativos, usando o léxico (F27). A *feature* F43, com a soma da polaridade dos termos no SenticNet, resultou em 5,836. Na *feature* F47 foram identificadas seis sentenças, separadas por uma barra vertical vermelha. Por fim, a *feature* F58 indicou que todas as 83 palavras do comentário estavam escritas corretamente.

Embora o conjunto final tenha 58 *features*, observa-se que em comentários curtos, muitos valores são zero, criando uma matriz de atributos esparsa. Isso é comum em contagens textuais e influenciou a escolha de algoritmos que lidam bem com dados esparsos.

3.3. Classificação dos dados por modelos de Aprendizagem de Máquina

Os modelos escolhidos foram: SVM, *Random Forest* (RF), *Logistic Regression* (LR), *Gradient Boosting Trees* (GBT) e XGB, por serem amplamente utilizados para problemas de classificação [de Oliveira and de Melo 2021, de Almeida Neto and de Melo 2023, Hanić et al. 2024]. Para a implementação dos métodos em Python, foi utilizada, principalmente, a biblioteca Scikit-Learn. Os artefatos gerados na fase de extração de *features* textuais foram utilizados como entrada para os modelos em uma divisão estratificada por categoria e *rating* com validação cruzada de 5 *folds*.

3.4. Métricas

Como o problema de *rating prediction* também pode ser tratado como uma tarefa de regressão, foram adotadas as métricas RMSE e MAE, amplamente utilizadas nesta tarefa, para avaliar os resultados. Dessa forma, é possível analisar a tendência de erros dos modelos entre as cinco classes de *rating*, considerando a subjetividade inerente ao problema que pode causar confusão entre classes próximas na tarefa de 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 em termos das unidades originais dos dados, já a RMSE (1) enfatiza erros maiores, pois eleva as diferenças ao quadrado antes de

calcular a média, tornando-o mais sensível a *outliers* e, por isso, essa métrica foi adotada para otimização de hiperparâmetros. Em ambos os casos, quanto menores os valores, melhor o desempenho do modelo.

Além disso, a métrica AUC (*Area Under the Curve*) foi calculada para avaliar a capacidade do modelo de separar classes, medindo a área sob a curva ROC. A AUC varia de 0 a 1, com valores próximos de 1 indicando melhor desempenho. Classes 4 e 5 foram consideradas positivas e as demais negativas, transformando o problema em uma classificação binária. Assim, a AUC avalia a eficiência do modelo em distinguir classes, crucial para diferenciar avaliações como 1 e 5 estrelas, com maior consistência do que entre classes próximas [Kang et al. 2023].

4. Resultados

4.1. Comparação de classificadores para *rating prediction*

Para responder à PP1, foram avaliados cinco algoritmos sobre os mesmos 5 *folds* estratificados por categoria e *rating*. Cada modelo teve seus hiperparâmetros ajustados via busca aleatória para minimizar o RMSE, conforme protocolo disponível no repositório do trabalho. A Tabela 3 apresenta média, desvio-padrão e intervalo de confiança (IC) de MAE, RMSE e AUC. As diferenças foram testadas com Friedman (5 folds, α = 0,05) seguido do pós-hoc de Nemenyi.

Tabela 3. Resultados por modelo para cada métrica.

Grupo	↓ MAE	IC MAE	↓ RMSE	IC RMSE	↑ AUC	IC AUC
Gradient Boosting	0,9612 ± 0,0061	[0,9555, 0,9669]	1,3727 ± 0,0051	[1,3680, 1,3775]	0,7318 ± 0,0058	[0,7264, 0,7373]
Logistic Regression	1,0149 ± 0,0043	[1,0109, 1,0190]	1,4691 ± 0,0042	[1,4652, 1,4731]	0,7328 ± 0,0064	[0,7267, 0,7388]
Random Forest	0,9717 ± 0,0077	[0,9645, 0,9789]	1,4001 ± 0,0087	[1,3919, 1,4082]	0,7367 ± 0,0051	[0,7318, 0,7415]
XGBoost	**0,9499 ± 0,0047**	[0,9455, 0,9544]	**1,3721 ± 0,0055**	[1,3670, 1,3772]	**0,7390 ± 0,0046**	[0,7347, 0,7433]
SVM	1,0431 ± 0,0070	[1,0366, 1,0496]	1,5085 ± 0,0118	[1,4974, 1,5196]	0,7304 ± 0,0046	[0,7261, 0,7347]

O XGBoost obteve as menores MAE e RMSE e a maior AUC. O *pós-hoc* mostrou que ele supera Logistic Regression e SVM (MAE, RMSE) e não difere estatisticamente de Gradient Boosting ou Random Forest. Diante desse desempenho global e sua eficiência computacional, o XGBoost foi utilizado nos experimentos deste trabalho.

4.2. Comparação entre grupos de *features* textuais

A segunda pergunta de pesquisa (PP2) determina qual grupo de *features* é mais relevante para o problema de *rating prediction*. Para isso, utilizou-se o modelo XGB, que apresentou melhor desempenho na etapa anterior. Reaplicou-se o protocolo de otimização de hiperparâmetros, avaliando cada grupo de *features* individualmente. Os *folds* de validação cruzada continham apenas as *features* do grupo em análise. A Tabela 4 apresenta os resultados para cada grupo, com desvio padrão e intervalo de confiança para cada métrica.

Com 10 *features* ao todo, o grupo de léxico (LEX) obteve os melhores resultados em todas as métricas, indicando que informações sobre palavras negativas e positivas são cruciais em *rating prediction*. Esse resultado reflete a natureza subjetiva do problema, já que a polaridade das palavras indica a avaliação do usuário. Além disso, a quantidade de sentenças positivas ou negativas pode ajudar a classificar comentários com termos tanto positivos quanto negativos.

Tabela 4. Resultados por grupo de *features* para cada métrica.

Grupo	MAE	IC MAE	RMSE	IC RMSE	AUC	IC AUC
CONC	1,3160 ± 0,0108	[1,3059, 1,3261]	1,7589 ± 0,0131	[1,7466, 1,7712]	0,6243 ± 0,0025	[0,6220, 0,6267]
LEX	**1,1112 ± 0,0141**	**[1,0980, 1,1245]**	**1,5653 ± 0,0174**	**[1,5490, 1,5816]**	**0,6953 ± 0,0039**	**[0,6917, 0,6990]**
POS	1,2755 ± 0,0113	[1,2649, 1,2861]	1,7462 ± 0,0137	[1,7333, 1,7591]	0,6339 ± 0,0040	[0,6302, 0,6377]
EST	1,3905 ± 0,0047	[1,3862, 1,3949]	1,8446 ± 0,0068	[1,8382, 1,8510]	0,5936 ± 0,0036	[0,5903, 0,5970]
SYNT	1,5095 ± 0,0094	[1,5007, 1,5184]	2,0012 ± 0,0096	[1,9922, 2,0102]	0,6219 ± 0,0020	[0,6201, 0,6238]
TWT	1,5247 ± 0,0032	[1,5218, 1,5277]	2,0154 ± 0,0037	[2,0119, 2,0189]	0,6172 ± 0,0032	[0,6142, 0,6202]
MCL	1,4527 ± 0,0188	[1,4350, 1,4704]	1,9128 ± 0,0259	[1,8884, 1,9371]	0,5908 ± 0,0036	[0,5875, 0,5942]

Tabela 5. Resultados do *ablation study*.

Grupo	MAE	ΔMAE	RMSE	ΔRMSE	AUC	ΔAUC
Todos exceto CONC	0,9796 ± 0,0032	0,0297	1,4130 ± 0,0020	0,0409	0,7293 ± 0,0045	-0,0097
Todos exceto LEX	**1,0860 ± 0,0052**	**0,1361**	**1,5343 ± 0,0047**	**0,1622**	**0,6917 ± 0,0042**	**-0,0473**
Todos exceto POS	0,9889 ± 0,0032	0,0390	1,4121 ± 0,0052	0,0400	0,7305 ± 0,0033	-0,0085
Todos exceto EST	0,9575 ± 0,0034	0,0076	1,3830 ± 0,0045	0,0109	0,7389 ± 0,0044	-0,0001
Todos exceto SYNT	0,9616 ± 0,0054	0,0117	1,3872 ± 0,0088	0,0151	0,7350 ± 0,0034	-0,0040
Todos exceto TWT	0,9640 ± 0,0045	0,0141	1,3927 ± 0,0049	0,0206	0,7323 ± 0,0042	-0,0067
Todos exceto MCL	0,9535 ± 0,0041	0,0036	1,3798 ± 0,0060	0,0077	0,7385 ± 0,0038	-0,0005
Todos os grupos	0,9499	–	1,3721	–	0,7390	–

Após o grupo de léxico, os melhores resultados foram dos grupos POS e CONC. No grupo de POS, a quantidade de adjetivos, advérbios, pontuação, entre outros, é relevante para a análise do modelo. Da mesma forma, as *features* do grupo CONC carregavam informações relevantes sobre polaridade, como na *feature* de *pleasantness* (agradabilidade), que mede a emoção associada a determinada palavra. Contudo, o grupo de léxico utilizou um léxico específico para cada categoria do conjunto de dados, o que pode ter contribuído para seu desempenho superior.

Para aprofundar a análise, um *ablation study* foi conduzido com os grupos de *features*. Um *ablation study* treina o modelo com todos os grupos de *features*, exceto um, para avaliar como a remoção desse grupo afeta o desempenho do modelo e identificar qual grupo de *features* tem mais impacto. O modelo utilizado foi o XGB com validação cruzada, visando minimizar a métrica RMSE. A Tabela 5 mostra os resultados, com as colunas de "variação" representando a diferença entre o resultado do grupo e o obtido com todos os grupos de *features*, juntamente com os respectivos valores de desvio padrão. Os números em vermelho indicam que a métrica piorou após a remoção do grupo de *features*.

Os resultados indicam que todos os grupos de *features* contribuem para o problema. A remoção de qualquer grupo piora os resultados, sendo que o grupo léxico apresentou a maior perda. Portanto, assim como observado anteriormente, o grupo de *features* léxico é muito relevante para *rating prediction*.

4.3. *Recursive Feature Elimination*

A terceira pergunta de pesquisa (PP3) avalia o desempenho do melhor modelo em diferentes combinações de *features* textuais. Utilizou-se o algoritmo *Recursive Feature Elimination* para seleção de *features* [Chai et al. 2019]. O modelo foi treinado inicialmente com todas as *features* e, a cada iteração, a *feature* menos relevante é removida, e o modelo é re-treinado, repetindo até atingir o número mínimo de *features*. O treinamento foi feito com o modelo XGB, aplicando validação cruzada e os melhores hiperparâmetros

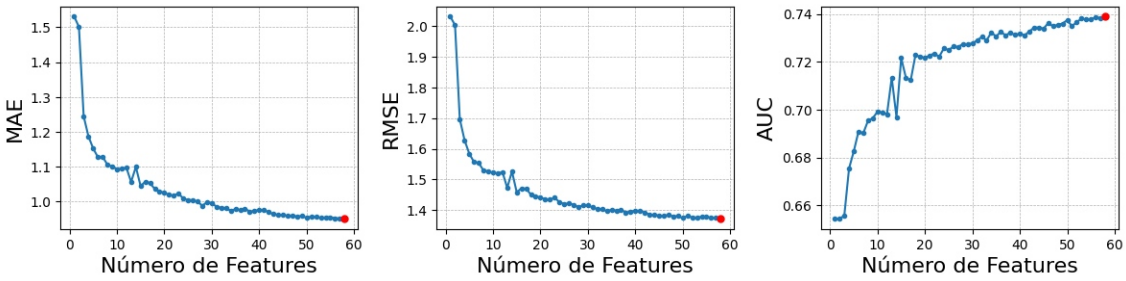

Figura 4. Gráficos de resultados do RFE para as métricas MAE, RMSE e AUC.

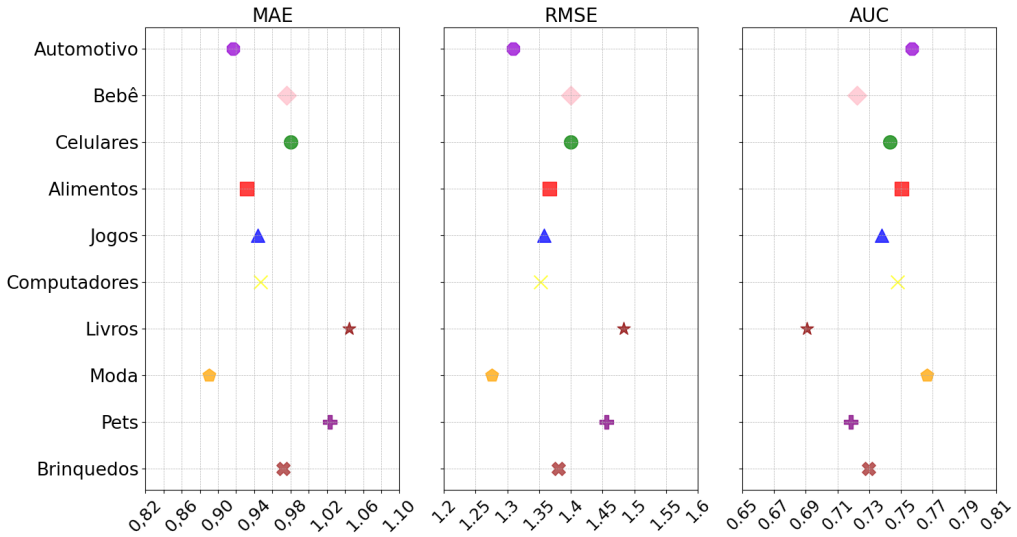

Figura 5. Resultados para diferentes categorias.

do primeiro experimento. A Figura 4 mostra três gráficos comparando as métricas MAE, RMSE e AUC em função do número de *features* selecionadas.

O melhor desempenho nas três métricas foi alcançado com o uso de todas as 58 *features*, indicando que nenhuma delas pôde ser descartada sem perda significativa de desempenho, reafirmando, assim, o resultado anterior do *ablation study*. Isso reforça a robustez do modelo XGB, capaz de capturar relações sutis mesmo entre variáveis menos relevantes, o que é especialmente importante em tarefas subjetivas como a predição de *ratings*. A *feature* mais relevante foi a contagem de palavras negativas do léxico (F27), evidenciando forte correlação com a variável de saída e confirmando achados anteriores.

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

A quarta pergunta de pesquisa (PP4) avalia se a configuração selecionada do modelo XGB mantém sua eficiência ao ser treinado em dados de diferentes categorias de produto. Para isso, o modelo foi ajustado utilizando validação cruzada e otimização por RMSE, com partições estratificadas por categoria. A Figura 5 mostra o desempenho por categoria, com cores e formatos de marcadores indicando cada categoria. A categoria *Moda* destacou-se, obtendo os melhores resultados em todas as métricas: MAE de 0,8899, RMSE de 1,2755 e AUC de 0,7663.

Moda é a segunda categoria com mais comentários, apenas atrás de *Livros*. No

entanto, *Livros* conta com 19.762 palavras únicas, incluindo 4.113 adjetivos únicos, enquanto *Moda* tem 6.394 palavras únicas e 1.275 adjetivos. Assim, *Livros* apresenta maior diversidade vocabular, pois diferentes gêneros literários podem ter critérios variados para avaliações, resultando em pior desempenho na classificação pela falta de padrões. Em contraste, a categoria de *Moda* tem um vocabulário mais homogêneo. Apesar da variedade de produtos, os consumidores costumam usar os mesmos termos para criticar ou elogiar, o que se reflete na menor quantidade de adjetivos únicos no subconjunto de dados. Essa homogeneidade facilita a identificação de padrões e melhora a classificação.

5. Conclusões

Este artigo abordou o problema de *rating prediction* com *features* textuais, usando um grande conjunto de avaliações da Amazon em português brasileiro. Avaliaram-se cinco modelos de aprendizagem de máquina, analisando o impacto das *features* em diferentes categorias. Os resultados mostraram que *features* de léxico são as mais relevantes, e o uso de *Recursive Feature Elimination* confirmou que todas contribuem em diferentes graus. A análise por categoria de produto mostrou que domínios com vocabulário homogêneo, como *Moda*, têm melhor desempenho, enquanto categorias com maior diversidade lexical, como *Livros*, apresentam desafios para a predição. Esses resultados enfatizam a importância da estrutura linguística do texto na modelagem de *rating prediction*.

Para trabalhos futuros, pretende-se explorar representações alternativas de texto e seu impacto na predição de *ratings*. Serão avaliadas técnicas clássicas, como TF-IDF e Bag of Words, e representações avançadas com *embeddings* de modelos de aprendizado profundo. Além disso, a inclusão de um conjunto de dados multilíngue permitirá investigar a capacidade de generalização dos modelos em diferentes idiomas e contextos.

Referências

A. Semary, N., Ahmed, W., Amin, K., Pławiak, P., and Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. *Plos one*, 19(2):e0294968.

Anchiêta, R. T., Neto, F. A. R., Marinho, J. C., do Nascimento, K. V., and Moura, R. S. (2021). Piln idpt 2021: Irony detection in portuguese texts with superficial features and embeddings. In *IberLEF@ SEPLN*, pages 917–924.

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.

Chai, Y., Lei, C., and Yin, C. (2019). Study on the influencing factors of online learning effect based on decision tree and recursive feature elimination. In *Proc. 10th Int. Conf. on E-Education, E-Business, E-Management and E-Learning (IC4E'19)*, pages 52–57.

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

Chenlo, J. M. and Losada, D. E. (2014). An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences*, 280:275–288.

de Almeida Neto, J. A. and de Melo, T. (2023). Exploring supervised learning models for multi-label text classification in brazilian restaurant reviews. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 126–140. SBC.

de Melo, T. (2021). Sentiprodbr: Building domain-specific sentiment lexicons for the portuguese language. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 349–354. SBC.

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.

de Oliveira, M. and de Melo, T. (2021). An empirical study of text features for identifying subjective sentences in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 374–388. Springer.

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.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., and Kaymak, U. (2015). Exploiting emoticons in polarity classification of text. *Journal of Web Engineering*, pages 022–040.

Hossain, M. I., Rahman, M., Ahmed, M. T., Rahman, M. S., and Islam, A. T. (2021). Rating prediction of product reviews of bangla language using machine learning algorithms. In *2021 International Conference on Artificial Intelligence 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*.

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.

Kimura, M. and Katsurai, M. (2018). Investigating the consistency of emoji sentiment lexicons constructed using different languages. In *Proc. 20th Int. Conf. on Information Integration and Web-based Applications & Services (iiWAS'18)*, pages 310–313.

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.

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

Stankevičius, L. and Lukoševičius, M. (2024). Extracting sentence embeddings from pretrained transformer models. *Applied Sciences*, 14(19):8887.