Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.8437

Research Paper

Rating Prediction in Brazilian Portuguese: From Classical Features to Large Language Models

Emanuelle Marreira [Universidade do Estado do Amazonas | *erm.eng22@uea.edu.br*]
Carlos Figueiredo [Universidade do Estado do Amazonas | *cfigueiredo@uea.edu.br*]
Tiago de Melo [Universidade do Estado do Amazonas | *tmelo@uea.edu.br*]

School of Technology, Universidade do Estado do Amazonas, 1200 Darcy Vargas Ave., Parque 10 de Novembro, Manaus, AM, 69050-020, Brazil.

Abstract. Online reviews play a crucial role in e-commerce, yet research on rating prediction for Brazilian Portuguese remains limited. This paper consolidates results from six interconnected studies investigating rating prediction and rating-text inconsistency detection. We evaluate approaches spanning classical machine learning with 58 textual features, BERT-based models, and ten large language models in zero-shot settings. Results show that BERTimbau achieves the best performance among fine-tuned models (MAE 0.56, RMSE 0.91), while DeepSeek and ChatGPT-4o lead among Large Language Models (LLMs) (RMSE 0.93). We also extend the analysis to a multilingual context with emoji signals across 13 European languages. For inconsistency detection, we find that LLM reliability varies substantially: ChatGPT-o3 shows low consistency across runs ($\kappa = 0.18$), while DeepSeek-3.2 achieves near-perfect agreement ($\kappa > 0.95$) with F1-score above 97%. Our findings provide practical guidelines for model selection based on accuracy requirements, training data availability, and cost constraints.

Keywords: Rating Prediction, Natural Language Processing, E-commerce

Received: 11 June 2026 • **Accepted:** 12 June 2026 • **Published:** 10 July 2026

1 Introduction

In the digital era, online platforms have become massive repositories of user-generated content, particularly textual reviews on e-commerce websites. These reviews, written by real consumers, serve as valuable resources that help others make well-informed purchasing decisions [de Melo *et al.*, 2019]. Within this context, automated sentiment extraction, known as sentiment analysis or opinion mining [Liu and Zhang, 2012], emerges as a powerful tool to derive actionable insights from such texts, supporting consumers, businesses, and recommendation systems alike [Li *et al.*, 2022].

Online reviews are typically accompanied by numerical ratings that indicate product quality from the consumer's perspective. These ratings commonly follow a star-based system, ranging from 1 to 5, where lower values express dissatisfaction and higher values indicate approval. Given their direct impact on product reputation and sales, the automatic inference of ratings from textual content, a task known as *rating prediction*, has become an important research problem [Ahmed and Ghabayen, 2022; Barman *et al.*, 2024]. This task is particularly relevant for platforms where explicit ratings are unavailable or unreliable, enabling the extraction of user sentiment at scale.

The evolution of Natural Language Processing (NLP) has introduced increasingly sophisticated approaches to rating prediction. Early methods relied on handcrafted textual features combined with classical machine learning (ML) models such as Support Vector Machines, Random Forests, and Gradient Boosting [Chambua and Niu, 2021]. The advent of deep learning brought contextual representations through models such as BERT, which significantly improved text classification performance [Devlin *et al.*, 2019; Gardazi *et al.*, 2025]. More recently, Large Language Models (LLMs) have demonstrated remarkable capabilities in zero-shot and few-shot settings, enabling rating prediction without task-specific training [Kang *et al.*, 2023]. Despite these advances, most existing studies focus on the English language, leaving a significant gap for other widely spoken languages.

Portuguese is one of the most used languages on the internet, yet research on rating prediction for Portuguese texts remains scarce [Pereira, 2021]. This gap is particularly notable for Brazilian Portuguese, which presents unique linguistic characteristics, including colloquialisms, regional expressions, and informal writing styles commonly found in online reviews. Furthermore, an underexplored challenge in this domain is the inconsistency between textual content and assigned ratings. For example, sometimes users write positive comments but assign low ratings, or vice versa. Such mismatches reduce the reliability of rating prediction systems and deserve dedicated investigation [Mudambi *et al.*, 2014; Almansour *et al.*, 2022].

This paper presents a systematic investigation of rating prediction approaches for Brazilian Portuguese reviews, consolidating results from six interconnected studies developed within the same research line. Our investigation spans the methodological spectrum from classical machine learning with textual features to state-of-the-art LLMs, while also addressing the problem of rating-text inconsistency. The main contributions of this work are:

- **Comprehensive benchmark:** We evaluate multiple approaches for rating prediction in Brazilian Portuguese, including classical ML models with 58 textual features [Marreira *et al.*, 2025c], BERT-based models (both Portuguese-specific and multilingual) [Marreira and

de Melo, 2025], and ten different LLMs in zero-shot settings [Marreira *et al*., 2025a].
- **Inconsistency detection:** We investigate the ability of LLMs to identify semantic mismatches between review text and assigned ratings, comparing model reliability and consistency across multiple experimental runs [Marreira *et al*., 2025b, 2026b].
- **Cross-lingual extension:** We extend our analysis to a multilingual context, examining emoji signals as predictive features across 13 European languages [Marreira *et al*., 2026a][1].
- **Integrated analysis:** We provide a comparative discussion of all approaches, highlighting trade-offs between accuracy, computational cost, and practical applicability.

Figure 1 illustrates the evolution of this research line, showing how each study built upon previous findings to progressively expand the scope and depth of investigation.

The remainder of this paper is organized as follows. Section 2 presents the methodology overview, describing the datasets and evaluation metrics used across studies. Section 3 details the rating prediction approaches, progressing from classical features to LLMs and multilingual emoji analysis. Section 4 addresses the detection of rating-text inconsistencies. Section 5 provides an integrated discussion comparing all approaches. Finally, Section 6 concludes the paper and outlines future directions.

2 Methodology Overview

This section provides an overview of the datasets and evaluation metrics used across the six studies, establishing a common ground for comparing results from different experimental setups.

2.1 Datasets

The studies in this research line rely on two main datasets collected by the authors: Amazon Brazil reviews for Portuguese-specific experiments and Google Play Store reviews for multilingual analysis. Table 1 summarizes the datasets used in each study.

The Amazon Brazil dataset was collected through web scraping from publicly available product reviews, covering ten product categories: Automotive, Baby, Cell Phones, Food, Games, Laptops, Books, Fashion, Pets, and Toys. Reviews include the textual content and a star rating from 1 to 5. The Portuguese-specific studies share this same source corpus but use different subsets according to each protocol. For the rating prediction studies, the dataset was balanced through undersampling to ensure equal representation across rating classes. For the inconsistency detection studies, only extreme ratings (1 and 5 stars) were considered, as these are more likely to exhibit clear sentiment polarity. The differences in dataset size across studies reflect the specific requirements and constraints of each experimental protocol. In particular, the LLM benchmark used a smaller test set because API-based experiments involve substantial request costs.

The Google Play Store dataset comprises reviews containing at least one emoji from 13 European locales (English, Spanish, Portuguese, French, German, Polish, Italian, Dutch, Romanian, Ukrainian, Czech, Greek, and Swedish). This dataset enables cross-lingual analysis of emoji usage patterns and their predictive capacity for ratings.

All collected review datasets are available through the corresponding original papers.

2.2 Evaluation Metrics

Given that rating prediction can be approached as both a classification and regression task, we adopt complementary metrics to assess model performance across studies.

For regression-oriented evaluation, we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE represents the average absolute difference between predicted and actual ratings, while RMSE penalizes larger errors more heavily. Lower values indicate better performance for both metrics.

For classification-oriented evaluation, we use F1-score (both micro and macro variants) and AUC (Area Under the ROC Curve). F1-micro treats each prediction equally regardless of class, while F1-macro computes the average F1 across classes, giving equal weight to each rating level. AUC measures the model's ability to distinguish between rating classes.

For the inconsistency detection studies, we additionally use Fleiss' kappa (κ), which generalizes agreement measurement to more than two raters to assess agreement across multiple independent executions of the same model. This metric is particularly important for assessing LLM reliability, as these models may produce different outputs for the same input across runs. Following the interpretation scale proposed by Landis and Koch [1977], κ values below 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and above 0.80 almost perfect agreement. Table 2 shows which metrics were used in each study, reflecting the specific objectives of each experimental setup.

3 Rating Prediction Approaches

3.1 Classical ML with Textual Features

The first study [Marreira *et al*., 2025c] investigated rating prediction using 58 handcrafted textual features combined with classical machine learning algorithms. The feature set was organized into seven categories: lexical features (number of positive words, number of negative words), syntactic features (number of occurrences of syntactic patterns such as an adjective followed by a noun), concept features (pleasantness average, polarity), part-of-speech features (number of nouns, number of adjectives), structural features (number of words, number of capital letters), Twitter/X features (number of elongated words, emojis) and miscellaneous features (number of words spelled correctly).

Five classifiers were evaluated: Support Vector Machines (SVM), Random Forest, Logistic Regression, Gradient Boosting Trees (GBT), and XGBoost. Experiments were conducted using 5-fold cross-validation on the balanced Amazon Brazil dataset.

The experimental results indicated that the XGBoost model achieved the highest overall predictive performance, with a MAE of 0.95, a RMSE of 1.37, and an AUC of 0.74. The feature importance analysis demonstrated that lexical

[1]The paper is under review.

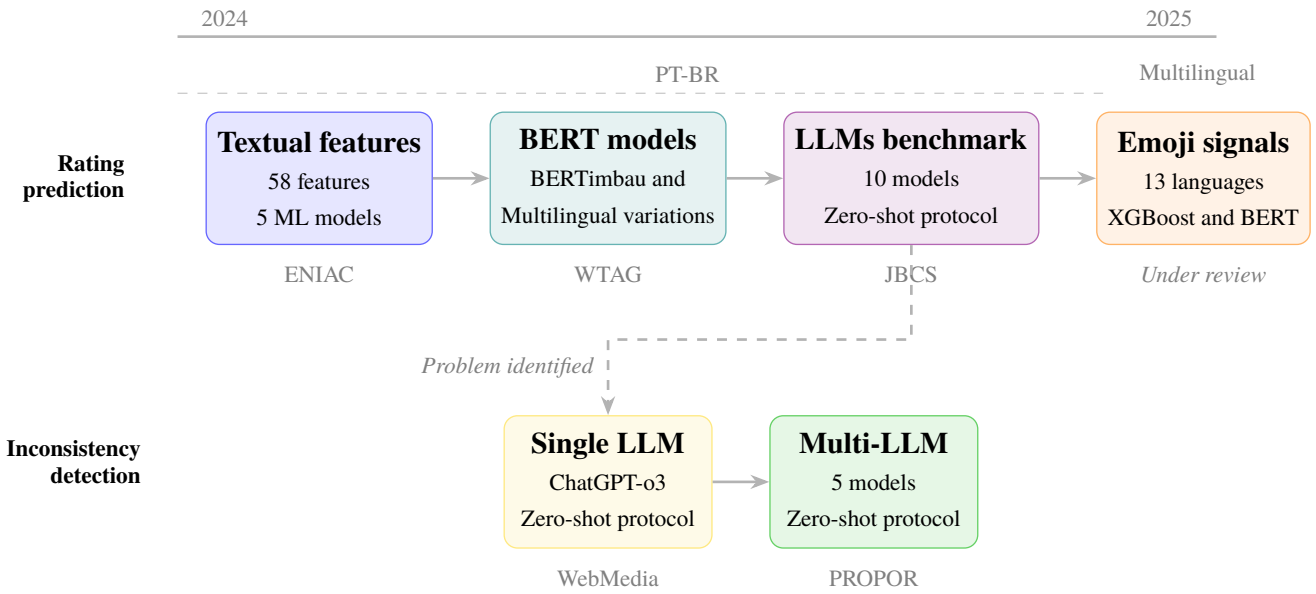

Figure 1. Evolution of the authors' published and under-review studies: from classical textual features to LLMs for rating prediction, and the emergence of inconsistency detection as a derived research problem.

Table 1. Summary of datasets used across studies.

Study	Source	Reviews	Categories	Period
Textual Features	Amazon BR	51,465	10	2021–2024
BERT Models	Amazon BR	51,465	10	2021–2024
LLMs Benchmark	Amazon BR	2,000	10	2021–2024
Inconsistency (single)	Amazon BR	20,586	10	2021–2024
Inconsistency (multi)	Amazon BR	20,586	10	2021–2024
Emoji Signals	Google Play	102,118	–	2012–2025

Table 2. Evaluation metrics used in each study.

Study	MAE	RMSE	AUC	F1	Fleiss' κ
Textual Features	✓	✓	✓	–	–
BERT Models	✓	✓	✓	–	–
LLMs Benchmark	–	✓	✓	✓	–
Inconsistency (single)	–	–	–	–	✓
Inconsistency (multi)	–	–	–	✓	✓
Emoji Signals	✓	–	–	✓	–

features, particularly indicators of polarity, constituted the most informative predictors. At the level of product categories, the Fashion domain yielded the most accurate predictions, whereas the Books category posed the greatest difficulty, a finding that is plausibly attributable to the greater length and higher semantic complexity of review texts in this domain.

This study is important because it validates interpretable automatic methods for mapping unstructured review texts into numerical ratings, helping consumers interpret large volumes of opinions and enabling companies to monitor customer satisfaction. It also shows that traditional machine learning models based on linguistic and lexical features can provide a viable and interpretable alternative to more complex approaches for Brazilian Portuguese reviews.

3.2 BERT-based Models

Building upon the limitations of handcrafted features, the second study [Marreira and de Melo, 2025] explored contextual embeddings through BERT-based models. Five variants were compared: BERTimbau (Portuguese-specific), multilingual BERT, DistilBERT, ALBERT, and RoBERTa.

BERTimbau significantly outperformed multilingual alternatives, achieving MAE of 0.56, RMSE of 0.91, and AUC of 0.93. This represents a substantial improvement over classical features: 41% reduction in MAE and 34% reduction in RMSE. The superiority of BERTimbau over multilingual BERT demonstrates the value of language-specific pre-training for Portuguese text understanding.

A category-wise analysis indicated that reviews in the Food and Baby domains were associated with the highest predictive accuracy, whereas those in the Laptop and Cell phones categories were substantially more challenging to model. Moreover, the performance disparities across categories were reduced relative to those obtained using classical feature representations, suggesting that contextual embeddings provide a more effective encoding of cross-domain semantic variability.

This study demonstrates the practical effectiveness of

Transformer-based architectures for mapping Amazon review texts to numerical ratings, with potential benefits for e-commerce applications and recommender systems. Its results also provide empirical evidence that, when fine-tuned for the task, a Portuguese-specific model can outperform general multilingual alternatives.

3.3 Large Language Models

The third study [Marreira *et al.*, 2025a] evaluated ten LLMs for zero-shot rating prediction, eliminating the need for task-specific training. The models included proprietary options (ChatGPT-3.5, ChatGPT-4o) and open-source alternatives (DeepSeek, Mistral, LLaMA 3, LLaMA 3.3, Gemma 1, Gemma 2, Sabiá-3, Sabiazinho). Two prompting strategies were compared: a simple prompt requesting only a rating and a detailed prompt explaining the rating scale.

ChatGPT-4o and DeepSeek achieved the best performance, with RMSE around 0.93, AUC of 0.91, and macro F1-score of approximately 0.53. Surprisingly, the simpler prompting strategy often yielded better results, suggesting that excessive instruction can introduce noise in model responses.

Cost analysis revealed significant differences: while ChatGPT-4o was the most expensive option, DeepSeek provided comparable performance at a fraction of the cost, making it an attractive alternative for large-scale applications. Portuguese-specific models (Sabiá-3, Sabiazinho) underperformed compared to multilingual LLMs, indicating that model scale and general capabilities currently outweigh language-specific training for this task.

An important finding was the relatively modest F1-scores (around 0.53) even for the best models, particularly for intermediate ratings (2, 3, and 4 stars). This suggests that rating prediction remains challenging even for state-of-the-art LLMs, motivating the investigation of rating-text inconsistencies discussed in Section 4.

By showing that zero-shot LLMs can be applied to online review rating prediction, this study highlights an alternative for organizations that need flexible rating inference without training and maintaining task-specific models. It also establishes a benchmark for Brazilian Portuguese, indicating which commercial, multilingual, and Portuguese-oriented models offer better trade-offs between predictive performance and cost.

3.4 Emoji Signals in Multilingual Context

The fourth study [Marreira *et al.*, 2026a] extended the analysis to a multilingual context, investigating emoji usage as a predictive signal for ratings. We collected 102,118 reviews from the Google Play Store across 13 European languages, including Portuguese, and examined whether emojis provide consistent cross-lingual information. This makes emoji-based signals a potentially useful complement to language-specific models, especially in multilingual settings where dedicated resources are not equally available for all languages. Thus, emoji signals may also provide useful cues in the Brazilian Portuguese context.

Three approaches were compared: text-only models, emoji-only models, and combined text+emoji models. For each of these three configurations, we evaluated two different classifiers: (i) a multilingual BERT model, and (ii) an XGBoost classifier trained on TF-IDF features. This setup allowed us to systematically compare the impact of input representation (text, emoji, or both) across both deep learning and traditional machine learning approaches.

Results showed that the combined BERT text+emoji model achieved the best performance with MAE of 0.87 and F1-micro of 0.41. Emoji-only models achieved surprisingly competitive results (approximately 70% of combined model performance), demonstrating that emojis carry substantial sentiment information. Cross-lingual analysis revealed that emoji sentiment patterns are relatively consistent across languages, with positive emojis (e.g., heart, thumbs-up) and negative emojis (e.g., angry-face) showing similar associations with ratings regardless of the review language.

These findings suggest that emojis can serve as lightweight, language-independent features for rating prediction. While text-based models remain more accurate, emoji signals offer a practical alternative for multilingual applications where language-specific models are unavailable or computationally prohibitive. Table 3 summarizes the main results across all rating prediction studies.

4 Rating-Text Inconsistency Detection

During the rating prediction experiments, we observed that some reviews exhibited clear mismatches between textual content and assigned ratings. For instance, enthusiastic praise accompanied by a 1-star rating, or harsh criticism paired with 5 stars. This phenomenon, which we term rating-text inconsistency, motivated two dedicated studies investigating whether LLMs can reliably detect such mismatches.

4.1 Initial Study with Single LLM

The first inconsistency study [Marreira *et al.*, 2025b] employed ChatGPT-o3 to classify reviews as coherent or incoherent. The experiment used 20,586 reviews with extreme ratings (1 or 5 stars) from the Amazon Brazil dataset. To assess model reliability, the same prompt was executed five times independently, and agreement across runs was measured using Fleiss' kappa.

Results revealed surprisingly low consistency. The global agreement across all five runs was $\kappa = 0.177$, indicating only slight agreement according to the Landis-Koch scale. Only 2% of reviews received unanimous classification across all runs, while 28.1% showed alignment with human judgment. Category-wise analysis showed that the Pets category exhibited the highest disagreement rate, suggesting that informal language and emotional expressions in pet-related reviews pose particular challenges for LLM interpretation.

These findings raise substantive concerns regarding the reliability of LLMs for subjective classification tasks. The pronounced variability observed across repeated runs, despite the use of identical prompts and inputs, indicates that reliance on a single-model configuration may be inadequate for achieving robust and reproducible inconsistency detection.

Motivated by recent advances in LLMs, this study serves as a critical warning against the unvalidated use of generative AI in review moderation workflows. It highlights the need for

Table 3. Summary of rating prediction results across studies.

Study	Best Model	MAE	RMSE	AUC
Textual Features	XGBoost	0.95	1.37	0.74
BERT Models	BERTimbau	0.56	0.91	0.93
LLMs Benchmark	DeepSeek/GPT-4o	–	0.93	0.91
Emoji Signals	BERT text+emoji	0.87	–	–

human validation and hybrid curation in commercial tasks that depend on subjective interpretation in e-commerce.

4.2 Multi-LLM Comparison

Based on the reliability issues observed in the initial study, the second investigation [Marreira *et al.*, 2026b] compared five LLMs: GPT-5, Llama-4, DeepSeek-3.2, Sabiá-3.1, and Bode-3.1. The same experimental protocol was applied, with three independent runs per model on the same 20,586 extreme-rating reviews.

DeepSeek-3.2 emerged as the most reliable model, achieving an F1-score above 97% for both coherent and incoherent classes, with Fleiss' $\kappa > 0.95$ indicating almost perfect agreement across runs. All five models demonstrated substantially higher consistency than ChatGPT-o3 from the previous study, with every model achieving $\kappa > 0.90$.

The proportion of reviews classified as incoherent varied across models, ranging from 8% to 12% of the corpus. Manual inspection of consistently flagged reviews confirmed that most represented genuine mismatches, such as reviews containing sarcasm, copy-paste errors, or apparent rating mistakes by users. Table 4 summarizes the inconsistency detection results.

In contrast to the initial single-model study, this comparison indicates that more recent LLMs may be operationally viable for identifying review inconsistencies. The results suggest that the textual content of some online reviews can contradict the star rating assigned by the user, which may mislead consumers. Therefore, such models can help platforms flag potentially misleading reviews, user input errors, or suspicious cases for further inspection.

5 Integrated Discussion

Regarding methodological progression, results demonstrate clear improvements as methods evolved. Classical textual features with XGBoost achieved an RMSE of 1.37, while BERTimbau reduced this to 0.91, resulting in a 34% improvement. Zero-shot LLMs achieved a comparable RMSE of 0.93 without any task-specific training, representing a different trade-off between performance and deployment simplicity. The emoji study showed that even simple signals can be informative, with emoji-only models achieving approximately 70% of the performance of text-based approaches.

The comparison between language-specific and multilingual models yielded nuanced insights. For fine-tuned models, BERTimbau clearly outperformed multilingual BERT. However, for zero-shot LLMs, multilingual models like DeepSeek and ChatGPT-4o outperformed Portuguese-specific alternatives like Sabiá. This suggests that model scale and general capabilities currently matter more than language specialization for zero-shot tasks, while language-specific pre-training remains valuable when fine-tuning is feasible.

Category-dependent performance was observed across all studies. Fashion and Food reviews were generally easier to predict, while Books, Games, and Laptops proved more challenging. This pattern likely reflects differences in review length, vocabulary complexity, and the degree of subjectivity in different product domains. Intermediate ratings (2, 3, and 4 stars) consistently posed the greatest difficulty, as the sentiment expressed in such reviews tends to be more nuanced and ambiguous.

The inconsistency detection studies revealed that approximately 10% of extreme-rating reviews contain mismatches between text and rating. More importantly, they highlighted significant reliability differences across LLMs. The contrast between ChatGPT-o3 ($\kappa = 0.18$) and DeepSeek-3.2 ($\kappa = 0.95$) demonstrates that model selection and prompt critically affect result consistency for subjective tasks.

From a practical standpoint, the studies suggest different recommendations depending on the use case. For applications requiring maximum accuracy with available training data, fine-tuned BERTimbau remains the best choice. For scenarios where training data is unavailable or deployment simplicity is prioritized, DeepSeek offers an attractive balance of performance and cost. For multilingual applications, emoji-based features provide a lightweight cross-lingual signal that can complement text-based models.

Several limitations should be acknowledged. All Portuguese studies used data from Amazon Brazil, which may not generalize to other e-commerce platforms or Portuguese varieties. The LLM benchmark used a sample of 2,000 reviews due to API costs, potentially affecting statistical power. The emoji study focused on European languages and may not capture emoji usage patterns in other regions.

6 Conclusion

This paper presents a systematic investigation of rating prediction approaches for Brazilian Portuguese reviews, consolidating results from six interconnected studies. The research progressed from classical machine learning with textual features through BERT-based models to large language models, while also addressing the emerging problem of rating-text inconsistency detection.

The main findings can be summarized as follows. First, contextual embeddings substantially outperform handcrafted features, with BERTimbau achieving 41% lower MAE than classical approaches. Second, zero-shot LLMs achieve competitive performance without training, with DeepSeek offering the best cost-performance trade-off. Third, emoji signals provide informative cross-lingual features that can enhance rating prediction in multilingual contexts. Fourth, LLM reliability varies dramatically for subjective tasks, emphasizing the need for careful model selection and validation.

Table 4. Summary of inconsistency detection results.

Study	Best Model	F1 (incoh.)	Fleiss' κ
Single LLM	ChatGPT-o3	–	0.18
Multi-LLM	DeepSeek-3.2	0.98	0.95

Taken together, these studies show that extracting quantitative information from unstructured online reviews is important for improving the reliability and efficiency of digital commerce ecosystems. Automated NLP techniques can transform subjective textual feedback into actionable knowledge, allowing companies to monitor public perception at scale, track long-term satisfaction trends, and identify potential user input errors or suspicious patterns. For consumers, these systems can reduce the effort required to interpret large volumes of feedback and support better-informed purchasing decisions. The results also indicate different deployment alternatives: LLMs can be reused for rating prediction without task-specific training, while simpler and less expensive models may still achieve competitive results in specific settings, such as those relying on emoji-based signals.

For future work, we identify several promising directions: investigating few-shot learning approaches to improve LLM performance on intermediate ratings; developing hybrid models that combine the interpretability of textual features with the power of contextual embeddings; extending inconsistency detection to identify specific types of mismatches (sarcasm, irony, errors, manipulation); and exploring the environmental impact of different modeling choices to promote sustainable NLP practices.

Declarations

Acknowledgements

The authors acknowledge the support provided by the Universidade do Estado do Amazonas (UEA) through the Academic Productivity Grant (GPA) (Administrative Ordinance No. 1177/2025-GR/UEA). This work was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

Authors' Contributions

All authors contributed equally to this work. They read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Further details on the datasets and their availability are provided in the corresponding original papers.

Further relevant information

The generative AI tool Gemini Pro was used strictly to support language translation and revision during the development of the article.

References

Ahmed, B. H. and Ghabayen, A. S. (2022). Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(7):3423–3432. DOI: 10.1007/s12652-020-01807-4.

Almansour, A., Alotaibi, R., and Alharbi, H. (2022). Text-rating review discrepancy (trrd): an integrative review and implications for research. *Future Business Journal*, 8(1):3. DOI: 10.1186/s43093-022-00114-y.

Barman, K. D., Bordoloi, B., Kumar, A., and Halder, A. (2024). Review rating predictions using improved deep learning architecture. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 468–472. IEEE. DOI: 10.1109/cicn63059.2024.10847509.

Chambua, J. and Niu, Z. (2021). Review text based rating prediction approaches: preference knowledge learning, representation and utilization. *Artificial Intelligence Review*, 54:1171–1200. DOI: 10.1007/s10462-020-09873-y.

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. DOI: 10.1016/j.ipm.2019.01.004.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the NAACL*, pages 4171–4186. DOI: 10.18653/v1/N19-1423.

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. DOI: 10.1007/s10462-025-11162-5.

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*. DOI: 10.48550/arXiv.2305.06474.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174. DOI: 10.2307/2529310.

Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., and Yu, Z. (2022). Text mining of user-generated content (ugc) for business applications in e-commerce: A systematic review. *Mathematics*, 10(19):3554. DOI: 10.3390/math10193554.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer. DOI: 10.1007/978-1-4614-3223-4_13.

Marreira, E. and de Melo, T. (2025). Predição numérica de avaliações em português: Comparando bertimbau e modelos multilíngues. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 57–63. SBC. DOI: 10.5753/sbbd_estendido.2025.247518.

Marreira, E., de Melo, T., de Oliveira, M., and Figueiredo, C. M. (2025a). Rating prediction in brazilian portuguese: A benchmark of large language models. *Journal of the Brazilian Computer Society*, 31(1):827–838. DOI: 10.5753/jbcs.2025.5667.

Marreira, E., de Melo, T., de Oliveira, M., and Figueiredo, C. M. S. (2026a). Emoji signals across europe: Rating prediction and cross-lingual analysis. *Natural Language Processing*. Under review.

Marreira, E., de Melo, T., Oliveira, M., and Maurício, C. (2025b). Detectando incoerências avaliativas em e-commerce com llms-um estudo de caso na amazon brasil. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 535–539. SBC. DOI: 10.5753/webmedia.2025.15952.

Marreira, E., de Oliveira, M., and de Melo, T. (2025c). Rating prediction in brazilian portuguese reviews: An approach based on textual features. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1352–1363. SBC. DOI: 10.5753/eniac.2025.11794.

Marreira, E., Figueiredo, C. M., and de Melo, T. (2026b). Rating–text mismatch in brazilian portuguese reviews: How reliable are zero-shot llms? In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pages 959–967.

Mudambi, S. M., Schuff, D., and Zhang, Z. (2014). Why aren't the stars aligned? an analysis of online review content and star ratings. In *2014 47th Hawaii International conference on system sciences*, pages 3139–3147. IEEE. DOI: 10.1109/HICSS.2014.389.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115. DOI: 10.1007/s10462-020-09870-1.