Sentimento vs. Nota: Detectando Discrepâncias em Avaliações de Aplicativos

Marcos Paulo Soares Moura Filho[1], **Rogério F. de Sousa**[1]

[1]Instituto Federal do Piauí - Campus Picos (IFPI)
Picos – PI – Brazil

marcos.paulo.s.m.filho@gmail.com, rogerio.sousa@ifpi.edu.br

Abstract. *Online reviews strongly influence consumer behavior, yet they often exhibit discrepancies between their numerical rating and the sentiment expressed in the text. This paper presents a quantitative analysis of these discrepancies using machine learning models applied to the UTL-Corpus, a corpus of Google Play Store reviews. Among the models tested for polarity classification, BERTimbau achieved the best performance (91% accuracy and 90% F1-score) and identified 1,772 discrepant reviews (17.7% of the total), most of which were positive. The study highlights the importance of analyzing textual content in addition to numerical ratings to enhance feedback interpretation and improve the effectiveness of recommendation systems.*

Resumo. *As avaliações online exercem forte influência no comportamento dos consumidores, mas costumam apresentar divergências entre a nota numérica e o sentimento expresso no texto. Este artigo apresenta uma análise quantitativa dessas discrepâncias utilizando modelos de aprendizado de máquina aplicados ao UTL-Corpus, um córpus de avaliações da Google Play Store. Dentre os modelos testados para classificação de polaridade, o BERTimbau obteve melhor desempenho (acurácia de 91% e F1-score de 90%) e identificou 1.772 avaliações discrepantes (17,7% do total), em sua maioria positivas. O estudo reforça a importância de analisar o conteúdo textual além das notas para aprimorar a compreensão de feedbacks e a eficácia dos sistemas de recomendação.*

1. Introdução

As avaliações online tornaram-se uma fonte essencial de informação para consumidores em decisões sobre produtos e serviços, especialmente em plataformas de comércio eletrônico, aplicativos, turismo e *streaming*. Comentários e notas atribuídos por usuários influenciam não apenas pela média das avaliações numéricas, mas também pelos aspectos subjetivos e emocionais expressos nos textos [Constantinides and Holleschovsky 2016, Chen et al. 2022, Ahn and Lee 2024].

Entretanto, é comum observar discrepâncias entre a nota atribuída e o sentimento presente no comentário textual [Almansour et al. 2022]. Essa incoerência compromete a confiabilidade das avaliações, prejudicando tanto a análise automatizada quanto a experiência do usuário, e reduzindo a credibilidade das plataformas [Shan et al. 2021, Almansour et al. 2022, Sadiq et al. 2021, Chang et al. 2024]. Sistemas de recomendação, por exemplo, que se baseiam principalmente em médias de notas, tornam-se vulneráveis

a essas inconsistências, resultando em decisões equivocadas por parte de consumidores e empresas [Xu et al. 2021, Fazzolari et al. 2017].

Nesse contexto, é necessário desenvolver abordagens capazes de detectar automaticamente avaliações discrepantes, garantindo maior fidelidade entre o conteúdo textual e a nota atribuída. A identificação dessas inconsistências pode aumentar a confiabilidade dos sistemas de recomendação, apoiar consumidores em decisões mais informadas e auxiliar empresas na interpretação de *feedbacks* reais. Além disso, tal coerência fortalece a transparência das plataformas e reduz o impacto de avaliações imprecisas ou enganosas.

Este trabalho propõe um modelo de aprendizado de máquina para identificação automática de avaliações discrepantes, utilizando como base o UTLCorpus [Sousa 2023], composto por avaliações de aplicativos da *Google Play Store*. O objetivo é avaliar a viabilidade do modelo em detectar essas inconsistências e propor sua futura integração em plataformas de avaliação. O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve o método proposto; a Seção 4 discute os resultados obtidos; e a Seção 5 traz as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

A discrepância entre avaliações textuais e notas numéricas em sistemas de *reviews* online tem sido amplamente discutida na literatura, evidenciando desafios para a confiabilidade e precisão desses sistemas. [Almansour et al. 2022] realizaram uma revisão integrativa sobre o fenômeno *Text-Rating Review Discrepancy* (TRRD), destacando que modelos baseados apenas em notas podem falhar na representação do real sentimento dos usuários, recomendando o uso de abordagens híbridas que combinem conteúdo textual e notas. Em uma análise de larga escala, [Fazzolari et al. 2017] estudaram 160.000 avaliações no *TripAdvisor* e observaram que as discrepâncias ocorrem com maior frequência em notas intermediárias, sugerindo uma presença comum de sentimentos mistos nesses casos. Já [Sadiq et al. 2021], ao investigarem avaliações da *Google Play Store*, identificaram que 25% das notas estavam enviesadas, muitas vezes inflacionadas, e que modelos baseados em *deep learning*, como CNNs, foram eficazes na previsão de notas com base no texto, alcançando até 94% de precisão.

Além da detecção de inconsistências, [Shan et al. 2021] exploraram como as discrepâncias podem ser utilizadas na identificação de avaliações falsas. O estudo evidenciou que *reviews* fraudulentos tendem a apresentar mais incoerências entre texto e nota, e, com isso, o modelo baseado em *Random Forest* atingiu 92,9% de precisão na detecção de falsificações. Esses achados reforçam a importância de considerar tanto aspectos qualitativos quanto quantitativos das avaliações para garantir maior transparência, confiabilidade e autenticidade em sistemas de recomendação, análise de sentimento e tomada de decisão automatizada.

3. Procedimentos Metodológicos

Este trabalho propõe a classificação da polaridade textual e a identificação e análise de discrepâncias entre o sentimento expresso e as notas atribuídas pelos usuários. Para isso, foram utilizados dois conjuntos de dados distintos, um para o treinamento do modelo e outro para sua aplicação. O processo foi dividido em quatro etapas principais, desde a seleção dos dados até a avaliação do desempenho do modelo.

3.1. Seleção e Processamento dos Dados

3.1.1. Fontes de Dados

Os dados utilizados neste estudo provêm de dois principais corpora:

- **TweetSentBR** [Brum and das Graças Volpe Nunes 2017]: Corpus de *tweets* em português brasileiro, manualmente anotado com polaridade de sentimentos (positivo, negativo e neutro). Esse conjunto de dados foi utilizado para treinar os modelos de análise de sentimentos, para que a classificação de sentimentos seja feita sem viés introduzido pelas notas numéricas.
- **UTLCorpus** [Sousa 2023]: Conjunto de avaliações de aplicativos extraídas da Google Play Store, contendo textos de avaliações e suas respectivas notas de 1 a 5 estrelas. Esse corpus foi utilizado para aplicar o modelo treinado no corpus TweetSentBR para análise de sentimentos e identificar discrepâncias entre os sentimentos textuais e as notas atribuídas.

3.1.2. Processamento dos Dados

Para garantir a qualidade dos dados utilizados nos modelos de aprendizado de máquina, foi realizada uma etapa de pré-processamento distinta para cada conjunto: TweetSentBR e UTLCorpus. No TweetSentBR, inicialmente, foram filtradas apenas as classes positivas e negativas, descartando avaliações neutras com o objetivo de simplificar a tarefa de classificação. Em seguida, aplicou-se um balanceamento das classes via *downsampling*, assegurando uma distribuição equitativa entre os sentimentos. Além disso, *tweets* duplicados foram removidos para evitar interferência nos resultados.

No UTLCorpus, adotaram-se outras estratégias específicas. Avaliações neutras (com nota três) foram excluídas, por não serem consideradas discrepantes no contexto da análise. Textos curtos, com menos de três palavras, também foram descartados por não oferecerem contexto suficiente para inferência de sentimento. Assim como no TweetSentBR, avaliações duplicadas foram removidas. Após esse processamento foi selecionado um subconjunto balanceado de 10.000 avaliações, composto por 2.500 exemplos de cada nota relevante ao estudo (1, 2, 4 e 5 estrelas), assegurando uma distribuição uniforme entre as categorias analisadas.Essas etapas garantem que o modelo opere com dados relevantes e balanceados, contribuindo para a eficácia na detecção de sentimentos e discrepâncias.

3.2. Arquitetura, Treinamento e Aplicação dos Modelos

Foram avaliados sete modelos baseados na arquitetura Transformer: BERT Base Cased, BERT Base Uncased, RoBERTa Base, DistilBERT, XLM-RoBERTa, BERTimbau e TabularisAI. Todos os modelos foram treinados sobre o corpus TweetSentBR [Brum and das Graças Volpe Nunes 2017], após uma etapa de processamento e balanceamento dos dados.

Os modelos foram avaliados utilizando métricas quantitativas como acurácia e F1-score, sendo o modelo com melhor desempenho selecionado para aplicação ao UTL-Corpus [Sousa 2023]. Esse modelo foi então aplicado ao subconjunto do UTL-Corpus citado anteriormente, onde as avaliações foram classificadas quanto ao sentimento. Por

fim, as discrepâncias entre o sentimento textual e a nota numérica foram identificadas e classificadas.

3.3. Identificação e Classificação das Discrepâncias

A discrepância entre o sentimento do texto e a nota atribuída foi classificada em três categorias:

- **Discrepância Positiva**: O texto expressa um sentimento negativo, mas a nota atribuída é alta (4 ou 5 estrelas).
- **Discrepância Negativa**: O texto expressa um sentimento positivo, mas a nota atribuída é baixa (1 ou 2 estrelas).
- **Avaliações Coerentes**: Quando o sentimento do texto está alinhado com a nota atribuída.

Figure 1. Exemplos de avaliações retiradas do UTLCorpus

A Figura 1 ilustra esses três casos com exemplos reais. No primeiro, o comentário sobre o Mozilla Firefox é positivo, mas a avaliação recebeu apenas uma estrela, caracterizando uma discrepância negativa. No segundo, o usuário critica o app *Dual Space*, mas ainda assim atribui cinco estrelas — uma discrepância positiva. Já no terceiro exemplo, o texto elogia o app *Candy Crush Saga* e o avalia com nota alta, representando uma avaliação coerente.

3.4. Validação Manual e Referência para Avaliação do Modelo

Para avaliar o desempenho do modelo de classificação de polaridade, foi necessária a construção de uma referência confiável, uma vez que o modelo foi treinado em um conjunto e testado em outro. Para isso, realizou-se uma anotação manual com quatro avaliadores humanos sobre uma amostra aleatória de 10% dos dados classificados automaticamente (1.000 textos). Cada anotador classificou os textos como *positivo*, *negativo* ou *não sei*, e os exemplos com duas ou mais respostas *não sei* foram descartados, resultando em 928 exemplos válidos. A concordância entre os anotadores foi considerada moderada, com coeficientes de *Fleiss' Kappa* e *Krippendorff's Alpha* ambos em 0,584.

A polaridade final de cada comentário foi definida por maioria simples (três ou mais anotadores em concordância), servindo como *ground truth* para a avaliação do modelo. Com essa referência, foi possível construir a matriz de confusão e calcular as métricas de desempenho, como acurácia, precisão, revocação e F1-score.

4. Resultados

Esta seção discute a avaliação do modelo de análise de sentimentos desenvolvido e examina a distribuição das discrepâncias nas avaliações online. O modelo foi avaliado com base nas métricas de F1-score e acurácia.

4.1. Desempenho dos Modelos de Transformers

O BERTimbau [Souza et al. 2020] destacou-se com o melhor desempenho, alcançando 89% tanto no F1-score quanto na acurácia, superando outros modelos como TabularisAI e XLMRoBerta, que ficaram com 83% e 87%, respectivamente. Esses resultados são representados na Tabela 1, que sintetiza os valores de F1-score e acurácia para cada modelo.

Table 1. Resultados dos modelos transformers

Modelo	Tempo	Acurácia	F1-score	Perda (aval.)	Perda (treino)
TabularisAI	7m17s	0,82	0,83	0,56	0,12
BERT Base Uncased	13m15s	0,83	0,83	0,57	0,13
BERT Base Cased	12m17s	0,82	0,82	0,51	0,17
DistilBERT	10m27s	0,81	0,82	0,48	0,12
RoBERTa Base	24m31s	0,80	0,80	0,51	0,22
XLM-RoBERTa	15m45s	0,87	0,87	0,38	0,16
BERTimbau	11m46s	0,89	0,89	0,45	0,04

4.2. Resultados da Avaliação Manual

Como explicado na metodologia (Subseção 3.4), os resultados da avaliação manual foram utilizados como referência para comparar as predições do modelo de classificação. A partir de uma amostra de 1.000 textos aleatórios avaliados por quatro anotadores humanos, 928 exemplos válidos foram considerados após a exclusão de amostras com duas ou mais respostas *não sei*. A polaridade de cada texto foi definida pela concordância de, no mínimo, três avaliadores.

A comparação entre as anotações humanas e as predições do modelo BERTimbau revelou um desempenho satisfatório, com um F1-score de 90% e uma acurácia de 91%. A Tabela 2 apresenta a matriz de confusão resultante, evidenciando 361 verdadeiros positivos e 486 verdadeiros negativos. No entanto, também foram registrados 57 falsos positivos e 24 falsos negativos.

Table 2. Matriz de Confusão

	Positivo Real	Negativo Real
Predito Positivo	361	57
Predito Negativo	24	486

A análise da matriz mostra que o modelo tem alta capacidade de identificar corretamente sentimentos positivos e negativos, mas apresenta limitações na distinção de nuances que levam a erros de classificação. O *recall* de 94% indica que o modelo consegue recuperar a maioria dos exemplos positivos, enquanto a precisão de 86% sugere que há margem para melhorias, principalmente na redução dos falsos positivos — ou seja, casos em que sentimentos negativos foram classificados como positivos. Esses resultados demonstram que, apesar da boa performance, o modelo ainda pode ser aprimorado, especialmente em contextos mais ambíguos ou com linguagem neutra.

4.3. Análise das Avaliações Discrepantes

O modelo treinado foi aplicado a um subconjunto do *UTLCorpus*, contendo 10.000 avaliações de aplicativos. A partir da classificação automática dos sentimentos e da

comparação com as notas atribuídas pelos usuários, foram identificadas 1.772 avaliações discrepantes, representando aproximadamente 17,7% do total.

As discrepâncias foram divididas em dois tipos: **positivas**, quando o sentimento expresso no texto era negativo, mas a nota atribuída era alta; e **negativas**, quando o sentimento era positivo, mas a nota atribuída era baixa. Os resultados revelaram uma assimetria significativa: 82% das discrepâncias eram positivas (1.455 casos), enquanto apenas 18% eram negativas (317 casos).

Esse desequilíbrio pode sugerir uma tendência de alguns usuários a atribuirem notas relativamente elevadas mesmo quando expressam insatisfação no texto. Tal comportamento pode estar relacionado a hábitos de avaliação, gratidão pelo serviço recebido ou outras motivações subjetivas.

Esses achados corroboram estudos anteriores que apontam para a existência de um *viés positivo* em sistemas de avaliação, reforçando a importância de considerar o conteúdo textual como uma fonte complementar — e, muitas vezes, mais precisa — às notas numéricas. A identificação dessas discrepâncias pode auxiliar plataformas digitais na interpretação mais fiel do *feedback* dos usuários, contribuindo para a melhoria de seus sistemas de recomendação e tomada de decisão.

5. Conclusões e Trabalhos Futuros

Este trabalho teve como objetivo aplicar e avaliar modelos de aprendizado de máquina para análise de sentimentos em avaliações de aplicativos, visando identificar discrepâncias entre os sentimentos expressos nos textos e as notas atribuídas. Diversos modelos baseados na arquitetura *Transformer* foram testados, sendo o BERTimbau o que apresentou melhor desempenho, com acurácia de 91% e F1-score de 90% na validação manual conduzida com quatro participantes humanos. A avaliação também indicou concordância moderada entre os anotadores (Fleiss $\kappa = 0{,}58$; Krippendorff's $\alpha = 0{,}58$), e a matriz de confusão evidenciou a eficácia do modelo, apesar da presença de alguns falsos positivos (57) e falsos negativos (24).

Ao aplicar o modelo treinado sobre um subconjunto do UTLCorpus, composto por 10.000 avaliações da Google Play Store, foram identificadas 1.772 avaliações discrepantes — cerca de 17,7% do total. Essas discrepâncias foram categorizadas em positivas (sentimento negativo com nota alta) e negativas (sentimento positivo com nota baixa), sendo que 82% dos casos foram do tipo positivo. Esse padrão sugere uma tendência de usuários atribuírem notas elevadas mesmo quando expressam descontentamento textual, possivelmente motivados por fatores subjetivos, como hábitos culturais, gratidão pelo serviço ou expectativas pessoais.

Esses achados reforçam a importância de integrar a análise textual como recurso complementar às notas numéricas em sistemas de avaliação. A identificação automatizada de inconsistências contribui para uma melhor interpretação dos *feedbacks*, aumentando a confiabilidade de sistemas de recomendação. Como trabalho futuro, pretende-se expandir essa análise para outros domínios, como turismo ou e-commerce, além de investigar estratégias que lidem com aspectos linguísticos mais complexos, como ironia e ambiguidade, e integrar os modelos propostos a sistemas reais de recomendação.

References

Ahn, Y. and Lee, J. (2024). The impact of online reviews on consumers' purchase intentions: Examining the social influence of online reviews, group similarity, and self-construal. *Journal of Theoretical and Applied Electronic Commerce Research*, 19:1060–1078.

Almansour, A., Al-Otaibi, R., and Alharbi, H. (2022). Text-rating review discrepancy (trrd): an integrative review and implications for research. *Future Business Journal*, 8.

Brum, H. B. and das Graças Volpe Nunes, M. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Chang, H.-L., Liu, Y.-L., Keng, C.-J., and Jiang, H.-L. (2024). Examining discrepancies between online product ratings and sentiments expressed in review contents. *Management Analytics and Social Insights*, 1(1):129–144.

Chen, T., Samaranayake, P., Cen, X., Qi, M., and Lan, Y.-C. (2022). The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study. *Frontiers in Psychology*, 13.

Constantinides, E. and Holleschovsky, N. (2016). Impact of online product reviews on purchasing decisions. pages 271–278.

Fazzolari, M., Cozza, V., Petrocchi, M., and Spognardi, A. (2017). A study on text-score disagreement in online reviews. *Cognitive Computation*.

Sadiq, S., Umer, M., Ullah, D. S., Mirjalili, S., Rupapara, V., and NAPPI, M. (2021). Discrepancy detection between actual user reviews and numeric ratings of google app store using deep learning. *Expert Systems with Applications*, 181:115111.

Shan, G., Zhou, L., and Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, 144:113513.

Sousa, R. F. d. (2023). *Classificação da utilidade de opiniões em português brasileiro*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Xu, Z., Zeng, H., and Ai, Q. (2021). Understanding the effectiveness of reviews in e-commerce top-n recommendation. *ICTIR '21: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*.