Avaliação de Sentimentos de Aplicativos: Uma Comparação entre Modelos de Linguagem de Grande Escala

Kalidsa B. de Oliveira[1], **Gabriel M. Lunardi**[1], **Williamson Silva**[2]

[1]Centro de Tecnologia - Universidade Federal de Santa Maria (UFSM)
Av. Roraima nº 1000 –97105-900 – Santa Maria – RS – Brasil

[2]Universidade Federal do Pampa (UNIPAMPA)
Av. Tiaraju, 810 – 97546-550 – Alegrete – RS – Brasil

kalidsa.oliveira@ecomp.ufsm.br,gabriel.lunardi@ufsm.br

williamsonsilva@unipampa.edu.br

Abstract. *The expansion of e-commerce in Brazil has driven the demand for more efficient strategies to understand consumer perspectives. In this context, sentiment analysis emerges as an essential tool for examining user opinions on products and services. This research evaluates the performance of different Large Language Models (LLMs) with low computational cost, in the task of sentiment analysis.*

Resumo. *A expansão do e-commerce no Brasil tem estimulado a demanda por estratégias mais eficientes para entender a visão dos consumidores. Nesse cenário, a análise de sentimentos se apresenta como um instrumento essencial para examinar opiniões de usuários acerca de produtos e serviços. Esta pesquisa avalia a performance de diferentes Modelos de Linguagem de Grande Escala (LLMs), de baixo custo computacional, na tarefa de avaliação de sentiments.*

1. Introdução

O aumento expressivo do comércio eletrônico no Brasil nos últimos anos é resultado da expansão e da popularização dos aplicativos móveis para a venda de produtos, além da alteração nos costumes dos consumidores. O uso intenso de aplicativos, não só no comércio eletrônico, gera uma quantidade de dados cada vez menos tratável por métodos tradicionais de análise de dados. Logo, encontrar maneiras de extrair informação útil desses dados tornou-se uma necessidade latente [Átilas Barros 2024]. Nesse sentido, a análise de sentimentos surge como um instrumento para compreender as percepções e expectativas dos clientes em relação aos produtos e serviços disponibilizados, pois a realização de uma análise manual de comentários se torna impraticável, como podemos verificar no trabalho de [Mendes et al. 2022] com 1.6 milhões de comentários analisados.

Este trabalho tem por objetivo realizar uma comparação entre diferentes modelos de linguagem para a previsão de sentimentos em comentários de aplicativos móveis. Este estudo se baseia em aplicações anteriores dos autores como em [Souto Moreira et al. 2023, Siqueira et al. 2024a] e tendo, como principal motivação a necessidade de automatizar a análise de sentimentos em grandes volumes de texto. Para tal, são utilizadas versões compactas, acessíveis e econômicas de LLMs. O artigo contém

a seguinte estrutura: a Seção 2 apresenta o conjunto de dados e os procedimentos de pré-processamento dos mesmos, na Seção 3 é detalhada a metodologia empregada, a descrição dos modelos avaliados e os critérios de comparação, na Seção 4 engloba os resultado obtidos com uma análise do desempenho dos modelos segundo diferentes métricas e finalizamos com a Seção 5 que apresenta as considerações finais destacando as principais conclusões do estudo e possíveis direções para pesquisas futuras.Os resultados obtidos evidenciaram o desempenho desses modelos ao adotar métricas de avaliação como acurácia, precisão, revocação e F1-score.

2. Conjunto de Dados

O conjunto de dados utilizado neste trabalho possui 3.000 comentários, em português, dos seguintes aplicativos: Shopee, SHEIN, TikTok Lite, Nubank, Instagram, Photo & File Detect, Whatsapp Messenger, Canva: Desenho Fotos e Vídeos, Capcut - Editor de Vídeos e Gov.br. O *dataset* foi construído a partir do modelo de emoções fundamentais de Ekman por [Siqueira et al. 2024a] que apresenta avaliações classificadas manualmente em sete emoções básicas — felicidade, surpresa, tristeza, neutro, medo, desgosto e raiva — por meio de um processo de anotação colaborativa validado por múltiplos avaliadores. Os dados podem ser baixados em [Siqueira et al. 2024b]. Essa base foi construída com o objetivo de apoiar o desenvolvimento de aplicações para análise da experiência do usuário (*UX*), permitindo estudos sobre a relação entre emoções e satisfação dos usuários. Os comentários coletados possuem subjetividades relacionadas aos sentimentos e palavras que não agregam na análise destes. Suas categorias apresentam desbalanceamento, e foi decidido manter esse aspecto para observar o comportamento das LLMs.

No pré-processamento, utilizou-se bibliotecas de Processamento de Linguagem Natural (PLN) em Python, área da inteligência artificial que permite o entendimento dos computadores com a linguagem humana, tais como nltk e *Spacy*, bem como a biblioteca Pandas para a manipulação dos dados. O procedimento de extração de texto envolveu as seguintes fases: conversão de emojis para texto em português; remoção de caracteres não latinos e de pontuações, exceto nas traduções dos emojis, posteriormente substituídas por espaço em branco; conversão das strings em letras minúsculas, remoção de *stopwords* e termos irrelevantes; aplicação da técnica *Lemmatizer* que modifica as palavras de acordo com seu contexto gramatical. A distribuição dos dados foi realizada de maneira estratificada, assegurando a preservação das distribuições dos rótulos, separando 80% dos dados para treinamento e 20% para teste.

3. Metodologia

Neste estudo, avaliou-se diferentes LLMs populares a fim de comparar sua performance na tarefa de análise de sentimentos e emoções. Foram analisados os seguintes modelos pré-treinados: neuralmind/bert-base-portuguese-cased, meta-llama/Llama-3.2-1B, microsoft/deberta-base-mnli, FacebookAI/roberta-large-mnli, distilbert/distilbert-base-multilingual-cased e microsoft/Multilingual-MiniLM-L12-H384. Todos os experimentos e resultados podem ser conferidos no repositório do GitHub[1]. Para realizar o ajuste fino *(fine-tuning)* e avaliar os resultados das métricas, iniciou-se com a tokenização, utilizando o *AutoTokenizer* da biblioteca *transformers*, configurado para comprimir e preencher sequências até o limite máximo de 128 *tokens*. Os modelos, importados através do

[1]https://github.com/Kalidsa/ERBD-2025-Kalidsa

AutoModelForSequenceClassification, foram adaptados ao número de classes presentes nos dados. Depois de converter as bases de treinamento e teste mencionadas anteriormente em *Dataset* do *HuggingFace*, foi possível aplicar o mapeamento da função de tokenização e assegurar a conformidade com o modelo.

O treinamento foi realizado com o uso da classe *Trainer* e *TrainingArguments*, tendo como hiperparâmetros principais a taxa de aprendizado, tamanho do *batch* para treinamento e avaliação, número de épocas, estratégia de avaliação e decaimento de peso com os valores, respectivamente, de 0.00002, 10, 3, epoch e 0.01. A seleção dos hiperparâmetros foi baseada na necessidade de balancear o desempenho do modelo com a capacidade computacional, utilizando a GPU A100 do Google Colab para o experimento e estudos. Os autores [Bergmann 2024] destacam que uma taxa de aprendizagem menor (que reduz a magnitude de cada atualização dos pesos do modelo) tem menos probabilidade de levar a um esquecimento catastrófico. Todos os modelos passaram pelos mesmos processos de pré-processamento e treinamento para assegurar uma comparação equitativa.

4. Resultados

O propósito de cada modelo era fornecer os melhores resultados de forma equilibrada entre as métricas, com ênfase na avaliação da métrica *F1-score*, devido à necessidade de equilíbrio entre precisão e revocação. Isso é importante para um conjunto de dados desbalanceado, pois possibilita avaliar o rendimento global do modelo.

As Tabelas 1 e 2 sumarizam os resultados com as métricas adotadas para a previsão da polaridade e do sentimento, respectivamente. Os números sugerem que o modelo BERTimbau apresentou o melhor rendimento geral, seguido pelo DistilBERT e pelo LLaMa-3.2-1B. O MiniL exibiu um rendimento inferior nas métricas de *recall* e *F1-score*, indicando que não é adequado para essa tarefa específica. Modelos como RoBERTa e DeBERTa também apresentaram performance intermediária, porém ficaram aquém dos modelos mais eficientes, os três melhores citados anteriormente, em termos de precisão e *recall*. Estes resultados sugerem que modelos como o BERTimbau, podem ser mais apropriados para analisar sentimentos e emoções no contexto dos dados analisados em português. Entretanto, é preciso destacar que essa análise deve ser considerada com restrição, pois modelos de linguagem maiores e um conjunto de dados mais balanceado poderiam gerar resultados diferentes.

Tabela 1. Resultados de Classificação para Sentimento

Sentimento	BERTimbau	miniL	RoBERTa	distilBert	DeBERTa	LLaMa-3.2-1B
Precision	**0,723**	0,700	0,621	0,662	0,627	0,646
Recall	**0,721**	0,501	0,600	0,645	0,622	0,638
F1-Score	**0,711**	0,378	0,586	0,631	0,604	0,640
Accuracy	**0,721**	0,501	0,600	0,645	0,622	0,638

Em relação ao artigo de [Souto Moreira et al. 2023], observa-se que as técnicas empregadas no pré-processamento e as escolhas de manter o desbalanceamento, neste estudo, impactam diretamente os resultados das métricas, reduzindo os valores. No entanto, os LLMs, mesmo pequenos (1 bilhão de parâmetros), se mostraram úteis na classificação

Tabela 2. Resultados de Classificação para Polaridade

Polaridade	BERTimbau	miniL	RoBERTa	distilBert	DeBERTa	LLaMa-3.2-1B
Precision	**0,877**	0,825	0,813	0,833	0,814	0,827
Recall	**0,874**	0,819	0,821	0,828	0,824	0,839
F1-Score	**0,868**	0,808	0,814	0,817	0,817	0,832
Accuracy	**0,874**	0,819	0,821	0,828	0,824	0,839

de sentimentos em cenários desbalanceados, demonstrando desempenho satisfatório especialmente nas métricas de precisão e F1-score. A proeminência co BERTimbau e DistilBERT reflete a capacidade desses modelos em capturar nuances semânticas em textos em português. Embora o MiniL tenha limitações, os resultados sugerem que versões compactas de LLMs podem ser eficazes, equilibrando performance e requisitos computacionais.

5. Considerações Finais

Este artigo avaliou a capacidade de classificação (previsão) de sentimentos e emoções de LLMs para uma aplicação de avaliação da experiência de usuário. Os resultados indicaram a superioridade do modelo BERTimbau. Entretanto, é preciso cautela, pois LLMs maiores podem levar a resultados diferentes sendo, portanto, uma limitação deste estudo em função do poder computacional disponível. Nesse sentido, prospecta-se, como trabalhos futuros, a avaliação de modelos maiores acessados via APIs.

Agradecimentos

Os autores agradecem ao CNPq pelo apoio dos projetos Universais 405973/2021-7 e 402086/2023-6, bem como à FAPERGS pelo projeto: ARD/ARC – processo 24/2551-0000645-1.

Referências

Bergmann, D. (2024). O que é ajuste fino? `https://www.ibm.com/br-pt/topics/fine-tuning`. (Accessed on 08/02/2025).

Mendes, K. C. P., Paucar, V. L., and Filho, R. N. D. C. (2022). Análise de sentimentos de tweets utilizando diferentes técnicas de deep learning. In *Anais da Escola Regional de Engenharia de Software (ERES)*.

Siqueira, V., Costa, R. H., Soares, T., Lunardi, G., and Silva, W. (2024a). Dataset anotado de sentimentos a partir de comentários de aplicativos móveis. In *Anais do VI Dataset Showcase Workshop*, pages 65–76, Porto Alegre, RS, Brasil. SBC.

Siqueira, V., Lunardi, G. M., Silva, W., Hentges Costa, R. L., and Tales, S. S. (2024b). A dataset of polarities and emotions from brazilian portuguese play store reviews. `https://doi.org/10.5281/zenodo.10823148`.(Accessed on 13/02/2025.).

Souto Moreira, L., Machado Lunardi, G., de Oliveira Ribeiro, M., Silva, W., and Paulo Basso, F. (2023). A study of algorithm-based detection of fake news in brazilian election: Is bert the best. *IEEE Latin America Transactions*, 21(8):897–903.

Átilas Barros (2024). Desvendando narrativas e representações: a ia aplicada à análise de dados qualitativos. *Educare et Sabere*, 2(2). (Accessed on 08/02/2025).