# Classificação de Sentimento em Reviews de Jogos no Steam: Um Fluxo Completo de Machine Learning

**Arthur W. Almeida, Luiz Fernando Rodrigues, Rodrigo M. Pedreira**

Instituto Mauá de Tecnologia (IMT)
Praça Mauá, 1 - Mauá, São Caetano do Sul - SP, 09580-900

***Abstract.*** *This study presents a Machine Learning pipeline for sentiment analysis on game reviews collected from the Steam platform. The process encompasses text preprocessing through to the training and evaluation of binary classification models. Various ML models were tested, including Logistic Regression, K-Nearest Neighbors, Random Forest, and Naive Bayes, and evaluated using metrics such as accuracy, precision, recall, F1-score, MCC, and AUC-ROC. The models were optimized with cross-validation, and their probabilistic predictions were calibrated. Among the evaluated models, KNeighborsClassifier and Multinomial Naive Bayes demonstrated the best performance. The pipeline includes an analysis of the most influential terms in positive and negative reviews, enabling a clearer interpretation of the factors contributing to sentiment predictions. This work provides a practical and replicable approach for sentiment analysis in the context of game reviews, with potential to be extended with more advanced NLP techniques in future studies.*

***Resumo.*** *Este estudo apresenta um pipeline de Machine Learning para análise de sentimento em reviews de jogos coletados na plataforma Steam. O processo abrange desde o pré-processamento de texto até o treinamento e avaliação de modelos de classificação binária. Foram testados vários modelos de ML, incluindo Regressão Logística, K-Nearest Neighbors, Random Forest e Naive Bayes, avaliados por métricas como acurácia, precisão, recall, F1-score, MCC e AUC-ROC. Os modelos foram otimizados usando validação cruzada e suas previsões probabilísticas calibradas. Entre os modelos avaliados, KNeighborsClassifier e Multinomial Naive Bayes demonstraram o melhor desempenho. O pipeline inclui uma análise dos termos mais influentes em reviews positivos e negativos, permitindo uma interpretação mais clara dos fatores que contribuem para as previsões de sentimento. Este trabalho oferece uma abordagem prática e replicável para a análise de sentimento no contexto de reviews de jogos, com potencial para ser ampliada com técnicas de NLP mais avançadas em estudos futuros.*

## 1. INTRODUÇÃO

A análise de sentimento em reviews de jogos pode proporcionar insights valiosos para desenvolvedores e estúdios de jogos. Este estudo apresenta um pipeline completo de Machine Learning (ML) voltado à análise de sentimento de reviews de usuários. Com o

uso de métricas robustas e modelos variados, o objetivo é criar uma classificação automática precisa e confiável para o sentimento das avaliações.

## 2. OBJETIVO

O principal objetivo é treinar, testar e comparar a eficácia de modelos de ML em prever o sentimento de reviews como positivo ou negativo. Para isso, o código implementa várias fases de pré-processamento e utiliza múltiplos modelos de classificação, avaliados com diversas métricas de desempenho.

## 3. METODOLOGIA

### 3.1 Coleta e Pré-processamento de Dados

A primeira etapa do fluxo de trabalho é a coleta e processamento das reviews dos jogos. A função pre_ml_task é a principal responsável por essa fase, que envolve as seguintes etapas:

1. **Criação e Carregamento do Objeto Jogo**: O objeto jogo é instanciado ou carregado a partir de um arquivo salvo para facilitar a reutilização de dados.
2. **Construção do DataFrame de Reviews**: A função verifica se os dados de reviews já foram processados anteriormente. Caso contrário, os dados são obtidos e tratados para remover valores nulos, adicionar labels, limpar o texto, tokenizar e lematizar os termos.
3. **Visualização dos Dados**: A distribuição dos labels é plotada e as palavras mais frequentes são listadas, permitindo uma análise preliminar do conjunto de dados. O uso da WordCloud exibe as palavras mais comuns e sua frequência, oferecendo uma visão geral dos termos predominantes nos reviews.

### 3.2 Treinamento dos Modelos

Após o pré-processamento, diversos modelos de classificação binária são treinados. O processo começa com a divisão dos dados em conjuntos de treino e teste. O código oferece suporte para treinar uma variedade de modelos, incluindo:

- Regressão Logística
- K-Nearest Neighbors (KNN), *metric=cosine*
- Random Forest
- Naive Bayes Bernoulli, Multinomial e Gaussian

O treinamento dos modelos é realizado na função train_new, e os modelos são otimizados com GridSearchCV para maximizar sua precisão com base em parâmetros específicos.

## 4. AVALIAÇÃO DOS MODELOS

A função show_metrics_task exibe uma ampla variedade de métricas de desempenho, fundamentais para avaliar a eficácia dos modelos. As métricas usadas incluem:

- **Acurácia**: A proporção de previsões corretas.
- **Precisão, Recall e F1-score**: Para uma análise detalhada de desempenho.

- **ROC AUC**: Mede a capacidade do modelo de separar as classes.
- **MCC (Matthews Correlation Coefficient)**: Avalia a qualidade geral do modelo, considerando todas as classes.
- **Log loss**: Qualidade do ajuste quanto a porcentagem de certeza.
- **LR+ e LR-**: Valores que medem a qualidade da precisão probabilística quanto a falsos positivos e falsos negativos, respectivamente.
- **Brier Score**: Mede a precisão das previsões probabilísticas.

Essas métricas permitem uma análise multifacetada dos modelos, com destaque para o MCC, que fornece uma visão equilibrada de performance mesmo em contextos de classes desbalanceadas. Destaca-se que as três ultimas são as menos relevantes, pois está-se mais interessado no resultado da predição do que nas suas probabilidades.

## 4.1. Curvas ROC e DET

As curvas ROC e DET (Detection Error Tradeoff) são geradas para cada modelo, permitindo uma análise visual de sua capacidade discriminativa e da relação entre taxas de verdadeiros e falsos positivos. A curva ROC mostra o trade-off entre sensibilidade e especificidade, enquanto a curva DET destaca os erros de detecção.

## 4.2 Calibração das Probabilidades

O gráfico de calibração exibe o quão bem as probabilidades previstas estão alinhadas com as frequências reais, avaliando a calibração dos modelos. A curva de confiabilidade é especialmente importante para tarefas onde o objetivo é prever probabilidades precisas, além da simples classificação, que não é o caso aqui, mas pode ser interessante observar esses valores na escolha do modelo.

## 4.3 Impacto das Palavras

Para identificar quais palavras mais influenciam as classificações, o código calcula e exibe os efeitos médios das features. Essa etapa utiliza um modelo de Regressão Logística com um vetorizador TF-IDF para extrair as palavras que têm mais impacto na previsão. Os resultados são visualizados por meio de um gráfico de barras horizontal, com as palavras-chave mais influentes para cada categoria de sentimento.

# 5. RESULTADOS

Após o treinamento e avaliação dos modelos, alguns deles se destacaram em termos de desempenho. No exemplo dos reviews do jogo "Trials Fusion", as principais métricas são apresentadas:

| Modelo | Acurácia | Precisão | Recall | F1-score | ROC AUC | MCC | LR+ | LR- | Log Loss | Brier Score |
|---|---|---|---|---|---|---|---|---|---|---|
| KNeighborsClassifier | 0.775 | 0.756 | 0.737 | 0.744 | 0.827 | 0.492 | 2.208 | 0.222 | 0.530 | 0.177 |
| MultinomialNB | 0.768 | 0.780 | 0.695 | 0.709 | 0.857 | 0.467 | 1.711 | 0.135 | 0.477 | 0.156 |
| LogisticRegression | 0.728 | 0.759 | 0.628 | 0.627 | 0.842 | 0.365 | 1.365 | 0.134 | 0.507 | 0.169 |
| Bernoullinb | 0.734 | 0.715 | 0.671 | 0.680 | 0.810 | 0.383 | 1.631 | 0.257 | 0.926 | 0.222 |
| Randomforestclf | 0.705 | 0.689 | 0.615 | 0.615 | 0.759 | 0.295 | 1.337 | 0.271 | 0.618 | 0.196 |
| Gaussiannb | 0.689 | 0.653 | 0.637 | 0.641 | 0.639 | 0.290 | 1.508 | 0.407 | 11.147 | 0.311 |

Esses resultados mostram que o modelo KNeighborsClassifier obteve o melhor desempenho em termos de acurácia, MCC, e Brier Score. A análise do impacto das palavras revelou que termos como "fun", "good" e "great" são os mais comuns entre reviews positivos.

## 6. CONCLUSÃO

Este estudo apresentou um pipeline completo de Machine Learning para análise de sentimento de reviews no Steam, abordando desde o pré-processamento até a avaliação detalhada de modelos. O uso de múltiplas métricas e técnicas de visualização permitiu uma avaliação abrangente do desempenho dos modelos e dos principais termos que influenciam as previsões.

Os resultados indicam que o uso de modelos como KNeighborsClassifier e Multinomial Naive Bayes são promissores para a classificação de sentimento em reviews de jogos, enquanto o pipeline apresentado pode ser adaptado para outras análises de NLP (Natural Language Processing) com pequenas modificações.

## 7. REFERÊNCIAS


https://github.com/Rogerio-mack/IMT_CD_2024/blob/main/README.md

https://realpython.com/sentiment-analysis-python/

https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/

https://medium.com/dc6f17246aba/video-game-review-analysis-3c7602184668

https://mariofilho.com/guia-completo-sobre-roc-auc-em-machine-learning/#diferen%C3%A7a-entre-roc-auc-e-log-loss

https://mariofilho.com/o-que-e-acuracia-em-machine-learning/

https://www.nltk.org/

https://docs.streamlit.io/

https://scikit-learn.org/