Análise de Sentimento em Avaliações de Jogos da *Steam*: Um Estudo Comparativo entre *RoBERTa*, *SVM* e *Naive Bayes*

Alex Benhard Ferreira[1], **Daniel Lichtnow**[1]

[1]Colégio Politécnico – Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS

abenhardferreira@gmail.com, daniel.lichtnow@ufsm.br

Abstract. *Sentiment analysis aims to categorize texts according to the polarity expressed, generally classifying them as positive or negative. In this work, the application domain is related to digital games. The RoBERTa model (Robustly Optimized BERT Pretraining Approach) and traditional approaches, such as Support Vector Machines (SVM) and Naive Bayes, were used. The objective is to evaluate the performance of each approach within the application domain, as well as the influence of data volume and pre-processing strategies, in order to assist in choosing the most appropriate model and scenario.*

Resumo. *A análise de sentimentos tem como objetivo a categorização de textos quanto à polaridade expressa, geralmente classificando-os como positivos ou negativos. Neste trabalho, o domínio de aplicação está relacionado a jogos digitais. Foram utilizados o modelo RoBERTa (Robustly Optimized BERT Pretraining Approach) e abordagens tradicionais, como Support Vector Machines (SVM) e Naive Bayes. O objetivo é avaliar o desempenho de cada abordagem dentro do domínio de aplicação, bem como a influência do volume de dados e das estratégias de pré-processamento, de modo a auxiliar na escolha do modelo e do cenário mais adequados.*

1. Introdução

A Análise de Sentimento é voltada para a identificação e categorização de opiniões expressas em textos, sendo frequentemente, mas não somente, modelada como um problema de classificação binária — positiva ou negativa [Liu 2012]. Inicialmente, os métodos utilizados baseavam-se em dicionários de polaridade, onde palavras eram associadas a sentimentos positivos ou negativos. No entanto, com o avanço da Inteligência Artificial e do aprendizado profundo, modelos como o *BERT* e suas variantes passaram a capturar nuances mais complexas da linguagem [Devlin et al. 2018].

Este trabalho tem como objetivo avaliar modelos de classificação automática de sentimentos aplicados a avaliações de usuários de jogos publicadas na plataforma *Steam*. Serão realizadas comparações com algoritmos tradicionais de aprendizado de máquina, como *Support Vector Machines* (SVM) e *Naive Bayes* e o modelo *RoBERTa*, em função do crescente uso de abordagens baseadas em aprendizado profundo. Além descrever a realização do processo de análise de sentimento no domínio dos jogos digitais, o trabalho também busca fornecer, por meio dos experimentos, indicativos sobre os impactos que fatores como o tamanho dos *datasets* e aplicação das técnicas de pré-processamento textual

geram em cada modelo de forma a auxiliar interessados na área na avaliação do modelo a ser utilizado.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta o referencial teórico e os trabalhos relacionados; a Seção 3 descreve os recursos computacionais, ferramentas e *datasets* usados e gerados; a Seção 4 descreve o processo de treinamento e *fine-tuning* dos modelos; a Seção 5 apresenta os experimentos realizados com seus resultados e avaliação e a Seção 6 apresentas as considerações finais.

2. Referêncial Teórico e Trabalhos Relacionados

Os modelos de classificação de textos tradicionais/clássicos, que são baseados em estatística (e.g. *SVM*, *Naive Bayes*), foram dominantes até 2010, quando modelos baseados em aprendizagem profunda passaram a ser mais usados, obtendo em muitos casos, maior acurácia que os modelos tradicionais [Li et al. 2022]. Sobre os modelos de apresendizagem profunda cabe destacar o modelo *BERT (Bidirectional Encoder Representations from Transformers)*, introduzido no artigo [Devlin et al. 2018], que é baseado na arquitetura *Transformer* [Vaswani et al. 2017]. Já o modelo *RoBERTa* [Liu et al. 2019] é uma variante do *BERT*, pré-treinada com um volume ampliado de dados.

Dentre alguns dos trabalhos que comparam o modelo *BERT/RoBERTa* e modelos tradicionais está [Garrido-Merchan et al. 2023], onde são usados diferentes *datasets* e realizados experimentos com textos em idiomas distintos (inglês, português e chinês), sendo a melhor acurácia obtida, em todos os casos, usando *BERT*. Experimentos similares são apresentados em [Sabiri et al. 2023], onde a acurácia obtida usando *BERT* foi de 0,97 e usando *LinearSVC* 0,98, sendo ressaltado que no modelo *BERT* seria possível incrementar a performance com maior facilidade do que nos modelos clássicos (e.g. aumentando o número de épocas no treinamento, algo que não foi feito). Outro estudo compara *BERT* com modelos tradicionais para classificar notícias como falsas ou verdadeiras, sendo obtidos resultados superiores para o *BERT* [Moreira et al. 2023]. Em [Liao et al. 2021] usando o modelo *RoBERTa* foram observados também resultados superiores em relação a outras abordagens.

Portando, em vários trabalhos a superioridade dos resultados usando modelos de aprendizagem profunda em relação aos modelos clássicos é demonstrada. Este trabalho irá avaliar se o mesmo ocorre no domínio escolhido (avaliação de jogos) e também fatores que influenciam, algo que nem sempre é destacado em outros trabalhos.

3. Materiais: Recursos Computacionais, Ferramentas e *Datasets*

Nesta seção, são apresentados os materiais (recursos computacionais, ferramentas e *datasets*) utlizados. Inicialmente é descrita a plataforma *Steam* a partir da qual foram obtidos os dados para os experimentos.

3.1. Contexto do Estudo: Plataforma *Steam*

Nos experimentos foram usados dados da *Steam* uma plataforma que nasceu para facilitar a distribuição de jogos [Zhao 2024]. Em 2011 foi criado o *Steam Reviews*, permitindo a cada usuário escrever avaliações que são apresentadas nas página do produto na loja.

Para a execução dos experimentos e desenvolvimento deste trabalho, foi utilizado um sistema computacional com a configuração listada na Tabela 1.

Tabela 1. Especificações do Sistema Utilizado

Componente	Especificação
Processador	AMD Ryzen 5 5600 (6 núcleos, 12 threads, 3.5GHz base/4.4GHz boost)
Memória RAM	2 * 16GB DDR4 (2667MHz, dual-channel)
Placa de Vídeo	NVIDIA RTX 4060 (8GB GDDR6, 3072 CUDA cores)
Armazenamento	Kingston SNV2S1000G NVMe PCIe SSD (1TB, 3500MB/s leitura)
Sistema Operacional	Windows 10 Pro 64-bit

As principais ferramentas utilizadas foram:

- **Python 3.9+**;
- **fastText (lid.176)**: Biblioteca utilizada para identificação automática de idioma, [1]
- **HuggingFace Transformers (v4.28+)**: para o modelo *RoBERTa*;
- **Scikit-learn (v1.2+)**: para implementação dos modelos tradicionais (*SVM* com *SVC* e *Naive Bayes* com *MultinomialNB*), vetorização TF-IDF (*TfidfVectorizer*) e cálculo de métricas de avaliação.
- **SpaCy**: para tarefas de pré-processamento de texto.
- **NLTK (*Natural Language Toolkit*)**: remoção de stopwords e operações de PLN.
- **PyTorch (v2.0+)**: para implementação e treinamento do modelo *RoBERTa*.
- **CUDA (v11.7+)**: Plataforma de computação paralela para aceleração em GPU. [2]

3.2. Coleta e Seleção de Dados

A base de dados utilizada neste trabalho foi obtida na plataforma *Kaggle* denominada *Steam Store Apps Reviews* [3] base de dados que reúne aproximadamente 50 milhões de avaliações realizadas por usuários da plataforma *Steam*, contemplando diversos títulos de jogos e idiomas. Os dados estão disponibilizados em formato *JSON*. Optou-se por considerar apenas as avaliações redigidas em inglês. Essa decisão foi motivada, principalmente, por dois fatores: (1) a predominância desse idioma no conjunto de dados, o que viabiliza a obtenção de um número significativo de amostras e (2) a compatibilidade com modelos de linguagem, como o *RoBERTa* [Liu et al. 2019], que foi treinado majoritariamente com corpora em inglês.

Para preparar os dados para análise, foi desenvolvido um programa em Python responsável por filtrar apenas as avaliações em inglês (identificadas pelo campo *language: english*) e extrair os seguintes atributos considerados relevantes:

- ***review***: texto da avaliação escrita pelo usuário;
- ***voted_up***: indica se a avaliação foi positiva (*True*) ou negativa (*False*);
- ***votes_up***: número de usuários que consideraram a avaliação útil;
- ***playtime_at_review***: tempo de jogo no momento da avaliação, em horas;
- ***playtime_forever***: tempo total de jogo acumulado pelo usuário no jogo avaliado.

[1] Facebook AI Research. "fastText."`https://fasttext.cc/`. Acesso em: 15 maio 2025.

[2] NVIDIA. "CUDA Toolkit Documentation."`https://developer.nvidia.com/cuda-toolkit`. Acesso em: 15 maio 2025.

[3] https://www.kaggle.com/datasets/souyama/steam-reviews

Como resultado, obteve-se um conjunto com 3.000.931 avaliações em inglês. Desse total, 2.544.993 classificadas como positivas e 455.938 como negativas. Nota-se um desbalanceamento considerável entre as classes, com predominância de avaliações positivas. Para mitigar esse problema dois subconjuntos com 450.000 avaliações cada foram gerados: um composto por avaliações positivas e outro por negativas. A partir desses subconjuntos, foi construído um terceiro conjunto de dados balanceado, contendo 900.000 avaliações organizadas/ordenadas de forma alternada entre amostras positivas e negativas.

A partir do conjunto balanceado contendo 900.000 avaliações foram geradas duas versões do *dataset*, sendo realizado um pré-processamento básico nas duas versões:

- Remoção de *emojis* e caracteres especiais;
- Verificação do idioma real do campo *review*, utilizando a ferramenta *FastText*. Caso o conteúdo não estivesse em inglês, a revisão era descartada;
- Remoção de avaliações muito curtas (com uma ou duas palavras);
- Eliminação de URLs e links externos;
- Conversão de todo o texto para letras minúsculas (normalização de caixa).

Em uma das versões do *dataset* foi feita a remoção de *stopwords* usando a biblioteca *spaCy*. Porém as palavras como *bad* e *no* foram mantidas devido ao seu valor semântico [Jurafsky and Martin 2024]. A esta versão foi ainda feito um processo de lematização com o uso da biblioteca *NLTK*. No caso do modelo *RoBERTa* foi usada a biblioteca *HuggingFace Transformers*. Na Tabela 2 constam as três versões de *datasets* com o seu numero total de amostras extraídos a partir do *dataset* original que possuia 3.000.931 avaliações. A versão 1 desbalanceada foi gerada mediante a extração sequencial de avaliações do *dataset* original que possuia 3.000.931 avaliações. Todos os conjuntos de dados foram armazenados no formato *.pkl - Python Pickle*.

Tabela 2. Versões do *dataset* utilizadas nos experimentos

Versão	Tipo de Pré-processamento	Nº de Amostras
Versão 1	Conjunto original desbalanceado, sem remoção de *stopwords* nem lematização	885.093
Versão 2	Conjunto balanceado, sem remoção de *stopwords* nem lematização	900.000
Versão 3	Conjunto balanceado com remoção de *stopwords* e aplicação de lematização	900.000

4. Treinamento e Fine-Tuning

O treinamento dos modelos foi conduzido de maneira uniforme, utilizando os mesmos dados de entrada exceto quando claramente explícito. O processo de ajuste fino (*fine-tuning*) do modelo *RoBERTa* foi realizado com a biblioteca *HuggingFace Transformers*, utilizando os seguintes hiperparâmetros:

- **Taxa de aprendizado** (*learning rate*): adotou-se o valor de 5×10^{-5}, dentro do intervalo validado por [Liu et al. 2019] para *fine-tuning* de modelos *RoBERTa* para conjuntos de dados de média dimensão;

- **Tamanho do lote** (*batch size*): Foram utilizados 16 exemplos por dispositivo, tanto no treinamento quanto na avaliação;
- **Acúmulo de gradientes** (*gradient accumulation steps*): Com *gradient accumulation steps* = 2, obteve-se um *batch size* efetivo de 32, melhorando a estabilidade do treinamento sem ultrapassar os limites de memória da GPU;
- **Número de épocas** (*epochs*): Utilizou-se 5 *epochs* para o cenário com *holdout*.

A utilização de **precisão mista** (*fp16 = True*) contribuiu para a otimização do uso da GPU, permitindo treinar modelos maiores com maior eficiência computacional. A estratégia de treinamento incluiu avaliações periódicas a cada 1.000 passos, com salvamento automático do modelo com melhor desempenho com base na acurácia de validação.

5. Experimentos

Os experimentos foram realizados de forma a comparar a performance do *RoBERTa* com duas abordagens tradicionais de aprendizado supervisionado, *Naive Bayes* e *SVM*, para classificar avaliações de usuários da plataforma *Steam* como positivas ou negativas.

Para avaliação dos modelos, foram adotadas métricas calculadas com as bibliotecas *Scikit-Learn* e *HuggingFace Transformers*. As métricas usadas foram **Acurácia**, **Precisão**, **Revocação**, **F1-Score** (média harmônica entre precisão e revocação) e **Kappa de Cohen**. Os modelos de classificação foram treinados por meio de uma divisão simples dos dados (*holdout*), separando-se o conjunto em dados de treinamento (80%) e de teste (20%).

Nos experimentos realizados, foram considerados os seguintes aspectos:

- Comparação do desempenho dos modelos *Naive Bayes*, *SVM* e *RoBERTa* em conjuntos de dados balanceados e não balanceados;
- Impacto da remoção de *stopwords* e da lematização no desempenho dos modelos;
- Análise do impacto do tamanho do *dataset* nos resultados.

5.1. Avaliação do Impacto do Balanceamento do Conjunto de Dados

Para analisar o impacto do balanceamento de classes no desempenho dos modelos de classificação, foram utilizadas 15.000 amostras, sendo que uma foi obtida da versão 2 do *dataset* que possui balanceamento e outra da versão 1 desbalanceada do *dataset* original. Nenhum dos conjuntos passou por remoção de *stopwords* ou lematização, de modo a isolar o efeito exclusivo do balanceamento.

A Tabela 3 apresenta as métricas obtidas pelos três modelos em cada cenário. Nota-se que o balanceamento teve efeitos distintos dependendo da arquitetura do modelo. Observando as métricas, percebe-se que os modelos obtiveram acurácias mais altas com o conjunto de dados não balanceado. No entanto, esse ganho ocorreu em detrimento de um desempenho desequilibrado entre as classes. O coeficiente Kappa, que mede a concordância além do acaso, caiu para o *SVM* (de **0,6986** para **0,5774**) e principalmente para o *Naive Bayes* (de **0,6658** para **0,1076**), indicando que, apesar de acertos em volume, houve forte viés para a classe majoritária.

5.2. Impacto da Remoção de *Stopwords* e Lematização nos Modelos

A diferença entre os conjuntos avaliados nesta seção está na aplicação adicional — em apenas um dos conjuntos — da remoção de *stopwords* (com exceção de palavras de

Tabela 3. Desempenho com e sem balanceamento de classes (15.000 amostras)

Modelo	Balanceado	Acurácia	Precisão	Revocação	F1-score	Kappa
SVM	Sim	0,8493	0,8480	0,8480	0,8480	0,6986
SVM	Não	0,8700	0,9564	0,8861	0,9199	0,5774
Naive Bayes	Sim	0,8330	0,8496	0,8056	0,8271	0,6658
Naive Bayes	Não	0,8527	0,8517	0,9992	0,9195	0,1076
RoBERTa	Sim	0,9123	0,9328	0,8870	0,9093	0,8246
RoBERTa	Não	0,9347	0,9634	0,9589	0,9611	0,7561

negação como *"no", "not", "don't"*, entre outras) e da lematização dos termos. O objetivo é analisar o impacto dessas operações adicionais sobre o desempenho dos modelos de classificação. Foram conduzidos dois experimentos utilizando o mesmo conjunto balanceado de 15.000 amostras, correspondente a versão 2 (sem remoção de *stopwords* e lematização) e a versão 3 (com essas etapas adicionais), conforme a Tabela 2.

Para avaliação, ambos os experimentos adotaram a técnica de divisão simples (*holdout*) com 80% dos dados para treinamento e 20% para validação. A Tabela 4 apresenta os resultados de desempenho dos modelos *SVM, Naive Bayes e RoBERTa* com e sem a aplicação de remoção de *stopwords* e lematização.

Tabela 4. Desempenho com e sem proc. adicional (remoção de stopwords + lematização)

Modelo	P.Adicional	Acurácia	Precisão	Revocação	F1-score	Kappa
SVM	Não	0.8493	0.8480	0.8480	0.8480	0.6986
SVM	Sim	0.8383	0.8437	0.8342	0.8389	0.6767
Naive Bayes	Não	0.8330	0.8496	0.8056	0.8271	0.6658
Naive Bayes	Sim	0.8320	0.8617	0.7946	0.8268	0.6642
RoBERTa	Não	0.9133	0.9501	0.8709	0.9088	0.8265
RoBERTa	Sim	0.8717	0.8571	0.8950	0.8756	0.7432

O experimento demonstra que o impacto do pré-processamento textual varia conforme o tipo de modelo. O *RoBERTa* foi negativamente impactado, sugerindo que esse tipo de modelo se beneficia do texto completo, inclusive de *stopwords*. O *SVM* teve uma pequena perda, mas seu desempenho geral permaneceu estável. Já o *Naive Bayes* mostrou-se praticamente indiferente ao pré-processamento, com oscilações marginais nas métricas. Com base nesses resultados, conclui-se que a remoção de *stopwords* e a lematização podem ser contraproducentes, especialmente para modelos como o *RoBERTa*, que dependem da estrutura semântica e contextual do texto. Esses achados corroboram os estudos de [Siino et al. 2024], que indicam que práticas tradicionais de pré-processamento afetam de forma distinta os modelos tradicionais e pré-treinados.

Considerando a influência do pré-processamento, foi feita uma análise entre os modelos que considera o melhor cenário de teste para cada abordagem. Assim, *SVM* e *Naive Bayes* foram avaliados com a versão 3 do *dataset*(com remoção de *stopwords* + lematização) e *RoBERTa*, com a versão 2 (sem remoção de *stopwords* e sem lematização) usando *holdout*. Os resultados são apresentados na Tabela 5, sendo que o *RoBERTa* obteve os melhores resultados.

Tabela 5. Resultados comparativos utilizando *Holdout* - melhor cenário para cada abordagem

Modelo	Acurácia	Precisão	Recall	F1-Score	Kappa
SVM	0,8380	0,8461	0,8275	0,8367	0,6760
Naive Bayes	0,8385	0,8736	0,7926	0,8312	0,6771
RoBERTa	0,9060	0,8981	0,9144	0,9062	0,8120

5.3. Análise do Impacto do Tamanho da Amostra no Desempenho dos Modelos

Foram realizados experimentos com diferentes tamanhos de amostra do *dataset* (1.000, 10.000, 50.000 e 120.000 registros) para verificar seu impacto no desempenho dos modelos de classificação. Utilizou-se o *dataset* versão 2 e a estratégia de divisão holdout, com 80% dos dados para treinamento e 20% para teste. Aqui são apresentados os resultados para as métricas Acurácia, F1-score e Kappa.

A Figura 1 apresenta a variação da acurácia dos modelos em relação ao número de amostras. Já a Figura 2 ilustra sua a variação do F1-score conforme o tamanho da amostra e a Figura 3 ilustra a variação do coeficiente de Kappa. A Tabela 6 resume os resultados do experimento.

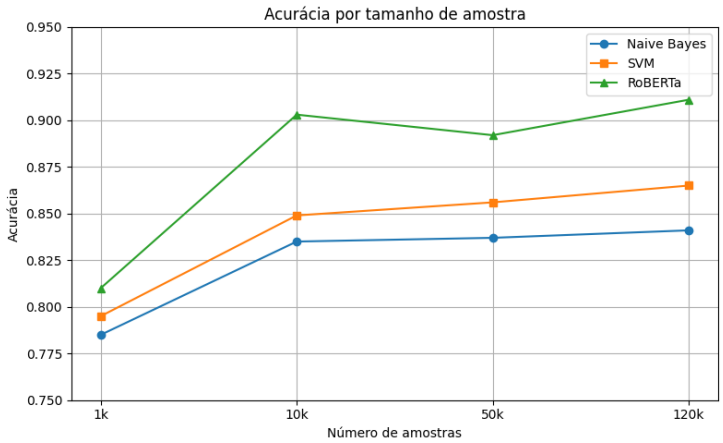

Figura 1. Evolução da acurácia dos modelos em função do tamanho da amostra

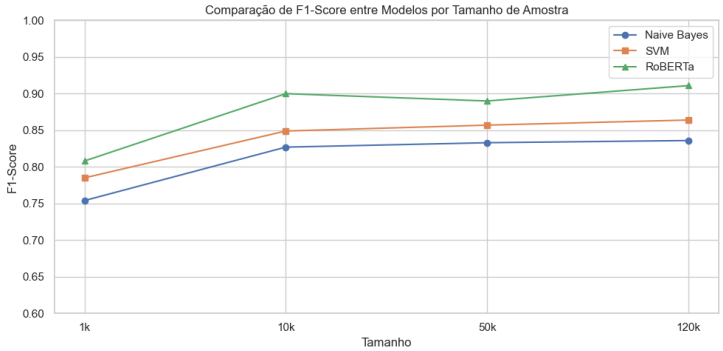

Figura 2. Evolução do F1-score dos modelos em função do tamanho da amostra

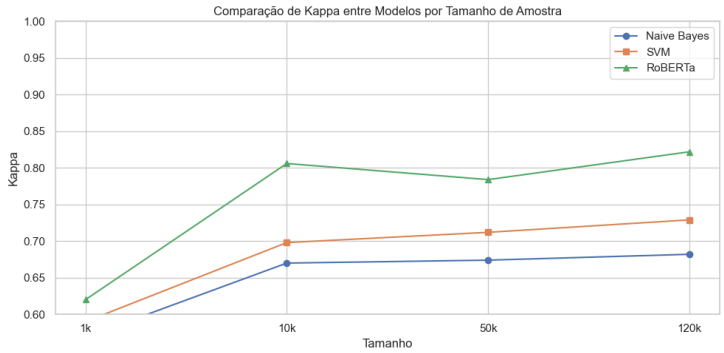

Figura 3. Evolução do coeficiente Kappa dos modelos em função do tamanho da amostra

Tabela 6. Resumo dos resultados para o tamanho da amostra

Modelo	**Métrica**	**1k**	**10k**	**50k**	**120k**
Naive Bayes	Acurácia	0,785	0,835	0,837	0,841
SVM	Acurácia	0,795	0,849	0,856	0,865
RoBERTa	Acurácia	0,810	0,903	0,892	0,911
Naive Bayes	F1-score	0,754	0,827	0,833	0,836
SVM	F1-score	0,785	0,849	0,857	0,864
RoBERTa	F1-score	0,808	0,900	0,890	0,911
Naive Bayes	Kappa	0,569	0,670	0,674	0,682
SVM	Kappa	0,590	0,698	0,712	0,729
RoBERTa	Kappa	0,620	0,806	0,784	0,822

A análise dos resultados evidenciou o impacto positivo do aumento da quantidade de amostras na performance dos modelos de classificação:

- O **RoBERTa** obteve os melhores resultados absolutos em todas as métricas com 120.000 amostras, incluindo acurácia de 91,1%, F1-score de 0,911 e Kappa de 0,822;
- Sobre o **SVM**, seu custo computacional consideravelmente mais baixo em comparação com modelos de aprendizagem profunda é um fator relevante: o treinamento com 120.000 amostras levou minutos, enquanto o RoBERTa demandou mais de 5 horas para a mesma tarefa;
- O **Naive Bayes** também se beneficiou do aumento de amostras.

É importante notar que pequenas variações, como a oscilação nas métricas do *RoBERTa* com 10k e com 50k amostras, não parecem invalidar a tendência geral de melhoria com maior volume de dados. Tais flutuações podem talvez ser explicadas pela variabilidade inerente à partição de teste na estratégia *holdout*.

Os experimentos indicam portanto, que o volume de dados exerce papel fundamental na performance dos modelos de aprendizado de máquina, especialmente naqueles com maior capacidade de aprendizado, como o *RoBERTa*. Em aplicações que exigem alta acurácia e confiabilidade, o investimento na coleta e utilização de grandes amostras,

aliado à escolha de modelos robustos, mostra-se vantajoso, ponderando-se o custo computacional associado.

A limitação de recursos computacionais representou um dos principais desafios. Enquanto os modelos *SVM* e *Naive Bayes* apresentaram baixo custo computacional e tempo reduzido de treinamento, o modelo *RoBERTa* exigiu considerável capacidade de processamento e um tempo significativamente maior para a realização dos treinamentos, devido à sua complexidade e à elevada quantidade de parâmetros envolvidos. A Tabela 7 apresenta os tempos aproximados de treinamento de cada modelo no experimento de avaliação do impacto do tamanho da amostra. Observa-se que os modelos *SVM* e *Naive Bayes* foram treinados utilizando exclusivamente o processador, enquanto o modelo *RoBERTa* foi executado com o uso de unidade de processamento gráfico (GPU). Cabe destacar que os tempos indicados são estimativas, visto que variações no uso dos recursos do sistema operacional podem impactar na duração dos treinamentos. Ressalta-se ainda que o uso de uma GPU compatível com *float 16* (FP16) no treinamento do *RoBERTa* resultou em uma redução de até 60% no tempo necessário para completar o processo.

Tabela 7. Tempo de treinamento aproximado dos modelos conforme o tamanho do conjunto de dados

Tamanho do conjunto	SVM + Naive Bayes	RoBERTa
1 mil	15 s a 30 s	˜1 min 30 s
10 mil	1 min a 1 min 30 s	˜15 min
50 mil	5 min a 7 min	1 h 30 min a 2 h
120 mil	20 min a 35 min	˜4 h 30 min a 5 h

6. Considerações Finais

Os resultados obtidos nos experimentos confirmam o ressaltado em diversos trabalhos citados na Seção 2 de que o modelos baseados em aprendizagem profunda (e.g. *RoBERTa*, *BERT*) vem tendo resultados melhores do que os tradicionais baseados em estatística para análise de sentimentos. Isto ocorreu também no caso do domínio de jogos digitais considerado neste trabalho. Mesmo chegando a este resultado, a realização de experimentos dentro de domínios específicos é importante, já que o domínio pode por vezes influenciar no acerto da categorização [Wahba et al. 2022]. Os resultados obtidos demonstraram a influência de fatores como tamanho do *dataset* e tipo de pré-processamento, análise pouco presente em trabalhos similares.

Dentre trabalhos futuros, está o aprofundamento da análise do impacto de técnicas de pré-processamento sobre os resultados, considerando o uso de modelos distintos. Outro trabalho futuro, dada a presença de sarcasmo e ironia em muitas avaliações da *Steam*, consiste em investigar e implementar técnicas específicas para a detecção dessas nuances linguísticas. Neste sentido, estudos indicam que o modelo *BERT* tem se mostrado superior a outros modelos [Eke et al. 2021] na identificação de sarcasmo. Outro trabalho futuro, dentro do domínio de avaliação de jogos, mas não limitado a ele, consiste em expandir a classificação binária (positivo/negativo) para incluir uma classe neutra ou para identificar emoções mais específicas (alegria, raiva, surpresa, etc.).

Referências

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Eke, C. I., Norman, A. A., and Shuib, L. (2021). Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model. *IEEE Access*, 9:48501–48518.

Garrido-Merchan, E. C., Gozalo-Brizuela, R., and Gonzalez-Carvajal, S. (2023). Comparing bert against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, 2(4):352–356.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. [s.n.], 3. ed. draft edition. Online manuscript released August 20, 2024.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

Liao, W., Zeng, B., Yin, X., and Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51(6):3522–3533.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, San Rafael, CA.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Moreira, L. S., Lunardi, G. M., de Oliveira Ribeiro, M., Silva, W., and Basso, F. P. (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.

Sabiri, B., Khtira, A., El Asri, B., and Rhanoui, M. (2023). Analyzing bert's performance compared to traditional text classification models. In *ICEIS (1)*, pages 572–582.

Siino, M., Tinnirello, I., and La Cascia, M. (2024). Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wahba, Y., Madhavji, N., and Steinbacher, J. (2022). A comparison of svm against pre-trained language models (plms) for text classification tasks. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 304–313. Springer.

Zhao, T. (2024). Valve corporation's evolution from a game developer to the gaming titan. *Transactions on Social Science, Education and Humanities Research*, 11:502–508.