INTELIGÊNCIA ARTIFICIAL E ÉTICA: UMA ANÁLISE COMPARATIVA ENTRE MÉTODOS DE DETECÇÃO DE TEXTOS GERADOS POR INTELIGÊNCIA ARTIFICIAL

Gilmar Valentim Camargo Júnior[1], Marlon de Matos de Oliveira[2]

Resumo: O avanço de grandes modelos de linguagem (LLMs), como ChatGPT e GPT-4, impulsionou um aumento nas pesquisas sobre detecção de conteúdo gerado por IA, motivado por preocupações éticas, acadêmicas e de desinformação. O trabalho de Wu et al. (2025), corrobora essa necessidade ao destacar os desafios crescentes da área, como as falhas de ferramentas comerciais, as limitações dos conjuntos de dados e a dificuldade humana em distinguir textos sintéticos. Inserido nesse contexto, este trabalho realiza uma avaliação comparativa de quatro classificadores: BERT, XGBoost, SVM e Regressão Logística, utilizando o dataset "AI vs Human Text". Após o treinamento com 60.000 textos, todos os modelos foram avaliados por métricas de acurácia, precisão, revocação e F1-score. Os resultados destacam o XGBoost com 99,26% de acurácia e o SVM com 99,23%, que, apesar do menor custo computacional em relação ao BERT, mostraram-se igualmente eficazes para a tarefa de detecção.

Palavras-chave: inteligência artificial; detecção de textos; BERT; XGBoost; SVM; GhostBuster; classificação de texto.

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), gilmarvcjr@gmail.com
[2]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), marlon.oliveira@unesc.net

ABSTRACT: The advancement of Large Language Models (LLMs), such as ChatGPT and GPT-4, has driven an increase in research on AI-generated content detection, motivated by ethical, academic, and disinformation concerns. The work by Wu et al. (2025) corroborates this need by highlighting the growing challenges in the field, such as the flaws of commercial tools, dataset limitations, and the human difficulty in distinguishing synthetic texts. In this context, this work conducts a comparative evaluation of four classifiers: BERT, XGBoost, SVM, and Logistic Regression, using the "AI vs Human Text"dataset. After training on 60.000 texts, all models were evaluated by accuracy, precision, recall, and F1-score metrics. The results highlight XGBoost with 99.26% accuracy and SVM with 99.23%, which, despite their lower computational costs compared to BERT, proved to be equally effective for the detection task.

Keywords: artificial intelligence; text detection; BERT; XGBoost; SVM; GhostBuster; text classification.

1 INTRODUÇÃO

A crescente evolução da Inteligência Artificial (IA), especialmente das IAs generativas de texto, tem transformado a forma como o conhecimento é produzido e disseminado. Essas tecnologias são capazes de gerar textos coesos e gramaticalmente corretos, muitas vezes difíceis de distinguir daqueles produzidos por seres humanos. Segundo Foletto (2023), a popularização de ferramentas como o ChatGPT provocou um intenso debate sobre os limites éticos e criativos do uso da IA na produção de conteúdo textual.

Com o uso massivo dessas ferramentas, surge uma problemática urgente: como assegurar a autenticidade de textos em contextos acadêmicos, educacionais e profissionais? A ausência de métodos padronizados e eficazes para detectar textos gerados por IA compromete a credibilidade da informação e levanta questões sobre autoria e propriedade intelectual. Ferro et al. (2019) afirmam que as criações de IA não são protegidas legalmente por falta de autoria humana direta, tornando-se de domínio público e levantando controvérsias sobre seu uso ético.

A escolha deste tema se justifica pela necessidade de oferecer mecanismos que identifiquem com precisão a origem textual, a fim de proteger direitos autorais e garantir a integridade acadêmica. Conforme (Uzun, 2023), a detecção eficaz de conteúdos gerados por IA é fundamental para

preservar a honestidade em ambientes educacionais, evitando a proliferação de plágio automatizado.

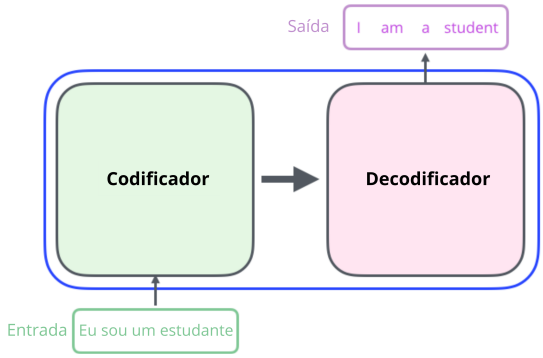
Pesquisas recentes têm se dedicado a avaliar métodos de detecção textual frente ao avanço das inteligências artificiais generativas. Shijaku e Canhasi (2023) demonstraram que modelos como o XGBoost são bastante eficazes na detecção de textos gerados por sistemas como o ChatGPT. De maneira semelhante, Cingillioglu (2023) apontaram o potencial dos modelos SVM para essa mesma tarefa, ressaltando sua eficiência mesmo com recursos computacionais limitados. Verma et al. (2024) propuseram o Ghostbuster, um classificador em múltiplas etapas que alcançou precisão acima de 99% na detecção de textos automatizados.

O objetivo geral desta pesquisa é **avaliar comparativamente métodos para a identificação de textos criados por modelos de inteligência artificial.** Para alcançar esse propósito, propõem-se como objetivos específicos a análise do panorama atual da Inteligência Artificial Generativa; a abordagem das implicações éticas e legais relacionadas à propriedade intelectual de conteúdos produzidos por IA; o treinamento de diferentes modelos de detecção já existentes; e, por fim, a avaliação do desempenho desses modelos a fim de identificar quais são os mais eficazes na distinção entre textos humanos e automatizados.

A detecção de textos gerados por Inteligência Artificial está diretamente relacionada ao avanço dos modelos de linguagem de grande porte, os chamados *Large Language Models* (LLMs). Esses modelos, como o GPT, são treinados em grandes volumes de dados textuais e têm a capacidade de aprender padrões complexos da linguagem humana, gerando textos coerentes e contextualmente relevantes. De acordo com Feuerriegel et al. (2024), os LLMs funcionam por meio da modelagem generativa, que simula a distribuição natural da linguagem com base em prompts de entrada, sendo capazes de produzir conteúdos que se assemelham muito ao discurso humano.

O surgimento da arquitetura *Transformer* em 2017, demonstrado na Figura 1, representou um ponto de virada na área. Diferentemente das Redes Neurais Recorrentes (RNNs), que processam as palavras de forma sequencial, os Transformers utilizam um mecanismo de atenção próprio, o self-attention, que permite analisar todas as partes relevantes de uma sequência de uma só vez. Essa abordagem garante maior compreensão contextual e dependências de longo alcance, contribuindo para a produção de textos mais fluentes e precisos (Gayam, 2022).

Figura 1 - Arquitetura Transformer



Fonte: Adaptado de Alammar e Grootendorst (2024)

Para alcançar os objetivos propostos, esta pesquisa fez uso de quatro métodos distintos de detecção de textos gerados por inteligência artificial: BERT, XGBoost, SVM e Regressão Logística. Esses modelos serão implementados, avaliados e comparados quanto à sua capacidade de classificar corretamente textos escritos por humanos e textos gerados por IA.

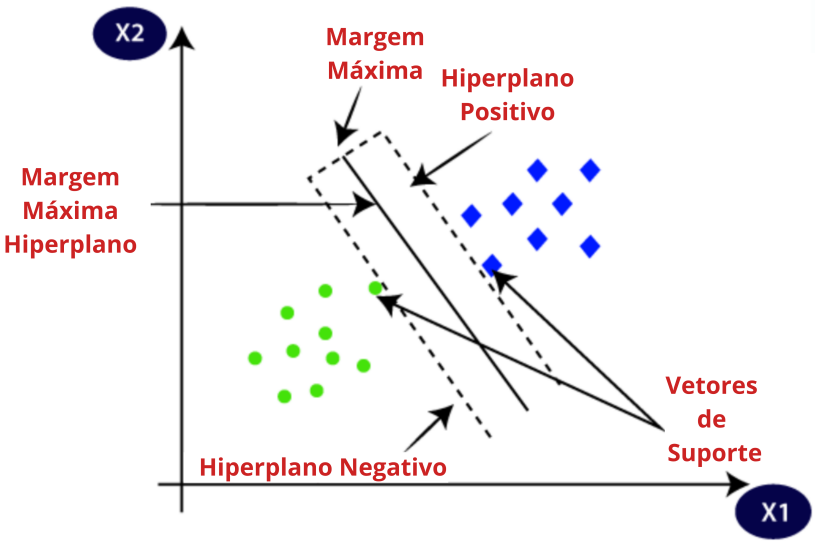
Bidirectional Encoder Representations from Transformers (BERT) é um modelo de linguagem pré-treinado baseado na arquitetura *Transformer*, desenvolvido para capturar o significado contextual das palavras em uma frase de forma bidirecional, ou seja, considerando tanto o que vem antes quanto o que vem depois de cada termo. Seu treinamento inicial é realizado com duas tarefas principais: o *Masked Language Modeling* (MLM), em que palavras são ocultadas e o modelo deve prever quais são, e o *Next Sentence Prediction* (NSP), que o ensina a entender relações entre sentenças consecutivas. Após esse pré-treinamento, o BERT pode ser ajustado à detecção de textos gerados por IA, adicionando uma camada de classificação supervisionada. Permitindo que o modelo aprenda a distinguir padrões sutis, como demonstrado no estudo de Wang, Li e Li (2024), em que o BERT alcançou resultados superiores a 97% de acurácia na tarefa de classificação binária entre textos humanos e automatizados.

O *Extreme Gradient Boosting* (XGBoost) é um algoritmo de aprendizado supervisionado baseado em árvores de decisão, que combina múltiplos classificadores simples para formar um modelo mais preciso por meio do ajuste sequencial de erros. No contexto da detecção de textos gerados por IA, ele atua classificando vetores numéricos extraídos dos textos, como representações TF-IDF ou características linguísticas específicas, como frequência de n-gramas ou uso de certas estruturas sintáticas. No estudo

de Shah et al. (2023), um modelo baseado em XGBoost foi capaz de identificar padrões recorrentes nos textos produzidos por IA, alcançando alta acurácia ao distinguir esses textos dos escritos por humanos.

O algoritmo *Support Vector Machine* (SVM) é eficaz na classificação de textos por conseguir lidar com dados de alta dimensionalidade e identificar fronteiras claras entre categorias. Aplicado à detecção de textos, ele aprende a separar textos de diferentes classes com base em padrões linguísticos. No contexto de identificação de textos gerados por IA, o SVM pode ser treinado com exemplos rotulados de textos humanos e artificiais. Máquinas de Vetores de Suporte procuram encontrar o hiperplano ótimo que melhor separa as classes em um conjunto de dados, assim como demonstrado na Figura 2. O termo hiperplano refere-se a uma fronteira de decisão que divide o espaço das características em duas regiões, uma para cada classe. As SVMs alcançam isso ao encontrar o hiperplano que maximiza a margem entre as classes. A margem é a distância entre o hiperplano e os pontos de dados mais próximos de ambas as classes, conhecidos como vetores de suporte (Sarkar et al., 2015).

Figura 2 - Separação de Classes com SVM



Fonte: Adaptado de Kapri (2020)

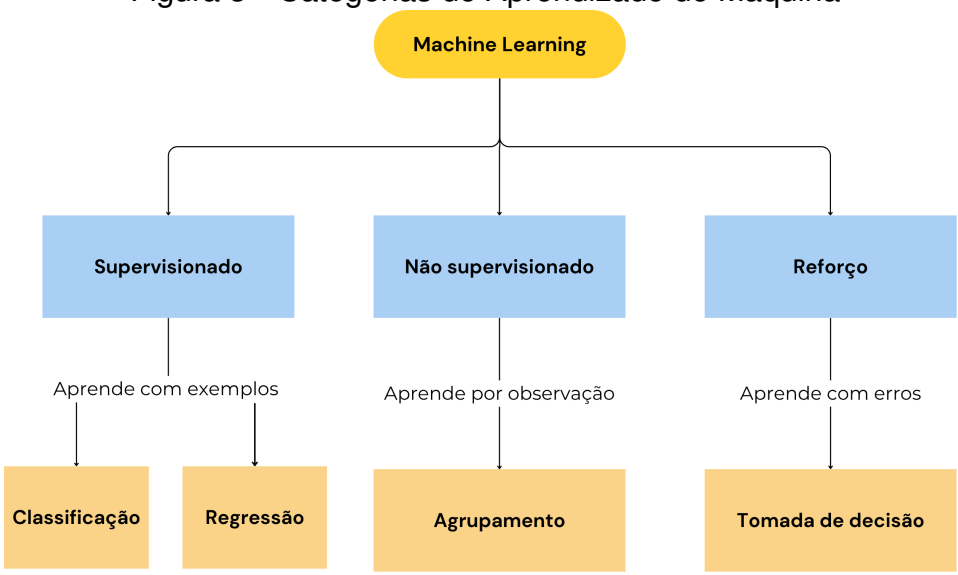
O *Ghostbuster* é um modelo avançado desenvolvido para a detecção de textos gerados por inteligência artificial, baseado em uma arquitetura de três etapas. Primeiramente, o sistema utiliza uma série de modelos de linguagem mais simples para calcular as probabilidades de ocorrência de cada token nos textos analisados. Em seguida, realiza uma busca es-

truturada para selecionar as características mais relevantes a partir dessas probabilidades, combinando vetores e escalas numéricas de forma estratégica. Por fim, essas características são utilizadas para treinar um classificador linear capaz de distinguir, com elevada precisão, textos gerados por IA de textos escritos por humanos. É considerado um método válido e altamente eficaz, tendo alcançado pontuação F1 superior a 99% em diversos conjuntos de dados, segundo seus desenvolvedores (Verma et al., 2024).

De acordo com Santos et al. (2005), a Regressão Logística é um modelo estatístico amplamente utilizado para a predição de variáveis categóricas binárias, ou seja, aquelas que assumem apenas dois possíveis resultados, como verdadeiro ou falso, 0 ou 1. No contexto deste trabalho, ela foi empregada para estimar a probabilidade de um texto ter sido gerado por inteligência artificial ou por um humano. Essa probabilidade está sempre compreendida entre 0 e 1 e é modelada em função de um conjunto de variáveis explicativas, geralmente obtidas por técnicas de vetorização textual como TF-IDF. A relação entre as variáveis independentes e a probabilidade do evento é estabelecida por meio de uma função logística, cuja forma sigmoide permite capturar padrões não-lineares entre os dados. Além da capacidade preditiva, a regressão logística também pode ser utilizada para análise de variáveis explicativas relevantes, auxiliando na interpretação e simplificação do modelo. Essa combinação de simplicidade, interpretabilidade e eficácia torna a regressão logística uma escolha apropriada para tarefas de classificação binária em textos.

Esses modelos são alimentados por técnicas de Aprendizado de Máquina (AM), conforme Figura 3, um subcampo da IA voltado para a criação de sistemas que aprendem a partir de dados. O AM pode ser dividido em três principais categorias: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, o modelo é treinado com exemplos rotulados, aprendendo a classificar novos dados com base nesses padrões. Já o aprendizado não supervisionado busca descobrir padrões ocultos em dados não rotulados, frequentemente por meio de agrupamentos. Por fim, o aprendizado por reforço envolve um agente que interage com o ambiente e aprende a tomar decisões com base em recompensas e penalizações (Mahesh, 2020).

Figura 3 - Categorias de Aprendizado de Máquina



Fonte: Elaborado pelo autor.

Diante da relevância do tema e da necessidade de soluções eficazes para a detecção de conteúdos automatizados, diversos estudos têm sido desenvolvidos no Brasil e no exterior, buscando compreender os impactos da IA na autoria textual e propor métodos de identificação.

O trabalho de Araujo (2016), intitulado “O uso de inteligência artificial para a geração automatizada de textos acadêmicos: plágio ou meta-autoria?”, investiga as implicações filosóficas e éticas da autoria quando algoritmos são capazes de produzir textos complexos e coerentes. O autor discute o conceito de "meta-autoria", exemplificando com casos reais, como o uso de bots para redigir notícias jornalísticas, e propõe uma reflexão sobre os limites da originalidade na era da automação.

No estudo de Almeida (2023) “Textos gerados por inteligência artificial e suas implicações no EaD”, analisa o uso crescente de ferramentas de IA por estudantes do ensino a distância, destacando a alta aceitação dos alunos em utilizar textos automatizados em atividades acadêmicas. O autor também descreve algumas ferramentas de detecção já disponíveis no mercado, mas aponta suas limitações, evidenciando a necessidade de soluções mais precisas.

Em “Inteligência Artificial (IA): o desafio para driblar o plágio do homem e agora da máquina”, Batistela et al. (2023) trata da evolução do plágio na era digital, abordando o uso de IA como ferramenta tanto de criação quanto de detecção de plágio. O artigo destaca a necessidade de políticas éticas, aprimoramento tecnológico e colaboração entre a comu-

nidade acadêmica e desenvolvedores para mitigar os riscos associados à autoria automatizada.

Cingillioglu (2023), em seu artigo “Detectando artigos gerados por IA: o desafio do ChatGPT”, apresenta um modelo de detecção baseado em Bag-of-Words com n-gramas, aplicado a textos gerados pelo ChatGPT. Utilizando algoritmos de aprendizado supervisionado, como SVM, o autor obteve alta precisão, com 92,7% de acerto, e nenhum falso positivo para textos humanos, o que demonstra a eficácia do modelo.

No artigo “Quem escreveu isto? Detectando texto gerado por inteligência artificial a partir de texto escrito por humanos” Kumar e Mindzak (2024) realizaram um estudo experimental com 135 participantes, solicitando que identificassem a origem de textos com base em sua leitura. Os resultados revelaram que apenas 24% dos textos gerados por IA foram corretamente identificados, apontando a dificuldade de detecção humana e a necessidade de métodos automatizados mais robustos.

Abdali et al. (2024) discutem, no artigo “Decodificando a caneta da IA: técnicas e desafios na detecção de texto gerado por inteligência artificial”, diferentes abordagens de detecção, como *watermarking*, modelos supervisionados, análise estilística e métodos baseados em recuperação. Os autores concluem que, apesar dos avanços, nenhuma técnica atual é totalmente confiável, especialmente diante de textos reescritos ou parafraseados.

Considerando os avanços recentes na geração de textos por inteligência artificial e os desafios associados à sua detecção, este trabalho se propõe a avaliar e comparar diferentes métodos de classificação para identificar textos gerados por IA, com base em uma base de dados rotulada e balanceada. A estrutura deste trabalho foi organizada para fornecer uma compreensão progressiva do tema. Em seguida, são apresentados os Materiais e Métodos da pesquisa, que descrevem os procedimentos utilizados para a preparação dos dados, os modelos implementados (BERT, XGBoost, SVM e Regressão Logística), e os critérios adotados para avaliação dos resultados. Na sequência, em Resultados e Discussões, são apresentadas as métricas obtidas com cada modelo, incluindo análises comparativas e visuais, como gráficos e matrizes de confusão, que ilustram o desempenho das abordagens. Por fim, a Conclusão retoma os objetivos propostos, discute as contribuições do estudo e propõe sugestões para pesquisas futuras.

2 MATERIAIS E MÉTODOS

A metodologia adotada neste trabalho se baseia na aplicação de algoritmos de aprendizado supervisionado e modelos de linguagem, como SVM, BERT, Regressão Logística e XGBoost, visando classificar textos como gerados por humanos ou por inteligência artificial. Para avaliar o desempenho dessas abordagens, são utilizadas métricas como acurácia, precisão, recall e F1-score, que permitem medir a eficácia dos classificadores em diferentes aspectos da tarefa. De acordo com Akram (2023), a análise dessas métricas é fundamental para compreender quais métodos são mais eficientes na detecção de conteúdo gerado por IA. O conjunto de dados utilizado é o "AI vs Human Text" Gerami (2024), composto por 487.235 textos rotulados, abrangendo exemplos criados tanto por pessoas quanto por modelos como GPT-3 e ChatGPT.

2.1 BASE DE DADOS

A base de dados utilizada nesta pesquisa é composta por 487.235 textos rotulados, sendo 305.797 gerados por inteligência artificial e 181.438 escritos por humanos. Os textos atribuídos à IA foram produzidos por modelos de linguagem como GPT-3 e ChatGPT, enquanto os textos humanos foram coletados de diversas fontes, como redações acadêmicas, fóruns públicos e outros registros autênticos. Para a implementação dos modelos de detecção, optou-se pelo uso de subconjuntos balanceados do conjunto original, contendo quantidades iguais de textos gerados por IA e por humanos, a fim de evitar viés de classe durante o treinamento e a avaliação dos algoritmos supervisionados empregados.

A escolha dessa base de dados se deu, principalmente, pelos seguintes motivos: sua escala, a curadoria clara das amostras e a aplicação direta ao problema investigado. O grande volume de textos disponíveis permite o treinamento e validação de modelos com boa representatividade estatística. Além disso, o conjunto já traz os textos rotulados como "IA" ou "humano", o que elimina a necessidade de rotulação manual e possibilita o uso direto de algoritmos supervisionados. Por fim, a presença de textos sobre diversos temas, como educação, sociedade, meio ambiente e tecnologia, permite que os modelos sejam avaliados em cenários variados, tornando os experimentos mais abrangentes e próximos da realidade de uso dessas ferramentas.

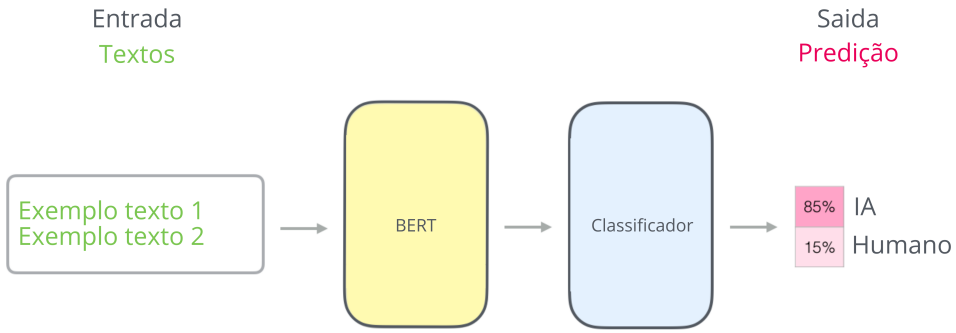
2.2 MÉTODOS DE DETECÇÃO

Para comparar a eficácia de diferentes algoritmos de aprendizado supervisionado na detecção de textos gerados por inteligência artificial, foram escolhidos quatro métodos amplamente utilizados para tarefas de classificação de texto: BERT, XGBoost, SVM e Regressão Logística.

2.2.1 Bidirectional Encoder Representations from Transformers

O modelo BERT representa o estado da arte em aprendizado profundo aplicado à linguagem natural, utilizando *embeddings* contextuais e pré-treinamento com textos de diversos idiomas em larga escala. Essa arquitetura, segundo Acheampong, Nunoo-Mensah e Chen (2021), representada pela Figura 4, permite a captura de relações semânticas complexas dentro do texto, indo além da simples frequência de termos, o que a torna particularmente eficiente para identificar padrões sutis em textos gerados por IA. Seu uso, entretanto, demanda maior capacidade computacional e tempo de treinamento mais elevado.

Figura 4 - Diagrama da arquitetura BERT.



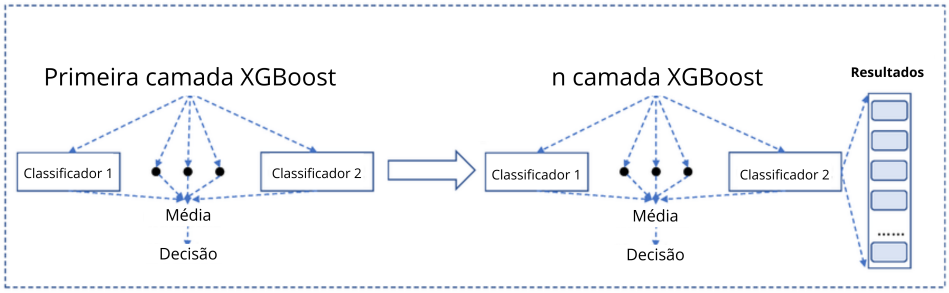
Fonte: Adaptado de Alammar (2018)

2.2.2 XGBoost

Segundo Ramraj et al. (2016), o XGBoost é um algoritmo que evoluiu do gradient boosting, utilizando árvores de decisão, especificamente árvores de regressão, como *weak learners* para fazer predições. Ele combina eficiência computacional com alta capacidade preditiva em sua implementação; árvores individuais são criadas usando múltiplos núcleos e os dados são organizados para minimizar os tempos de busca, o que diminui o tempo de treinamento dos modelos e, por sua vez, aumenta o desempenho. Essa abordagem, conforme a Figura 5, combina eficiência computacional e alto desempenho. Ao trabalhar sobre vetores TF-IDF derivados dos textos, o XGBoost consegue explorar interações não lineares entre os termos, o

que pode melhorar a discriminação entre textos humanos e artificiais. Esse método é conhecido pela robustez e pela capacidade de lidar com grandes volumes de dados de forma escalável.

Figura 5 - Fluxo de Decisão em Múltiplas Árvores no Modelo XGBoost.

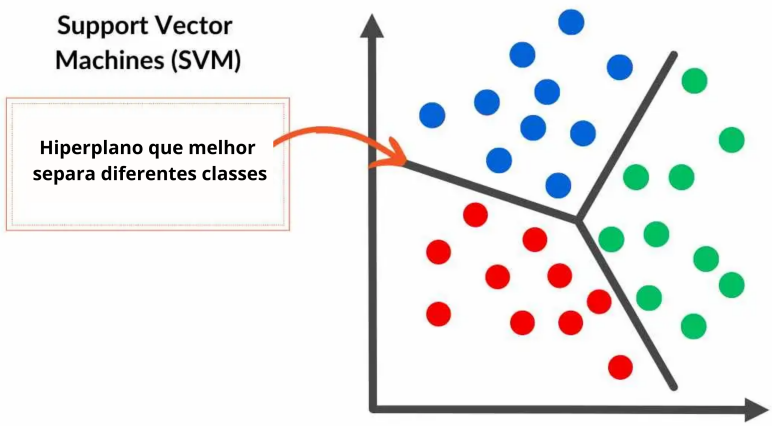


Fonte: Adaptado de Jiang et al. (2021)

2.2.3 SVM

De acordo com Chauhan, Dahiya e Sharma (2019), o SVM é uma técnica de classificação linear binária amplamente utilizada, especialmente eficaz em dados de alta dimensionalidade e tarefas de classificação binária, como os vetores de texto. Assim como demonstrado na Figura 6. A escolha de um kernel linear permite modelar limites de decisão simples, facilitando a interpretação dos resultados e mantendo a eficiência computacional. O SVM, ao maximizar a margem entre as classes, pode oferecer boa generalização, mesmo em contextos com ruído ou dados sobrepostos.

Figura 6 - Separação de classes com SVM.



Fonte: Adaptado de Otten (2024)

2.2.4 Regressão logística

Por fim, a Regressão Logística, conforme Miyazaki et al. (2024) é um método eficiente para classificação linear e frequentemente serve como um algoritmo de linha de base para comparação em Processamento de Linguagem Natural. Utilizando vetores de *features* derivados de texto como entrada (como contagens de palavras de léxico ou, potencialmente, TF-IDF), este modelo, devido à sua interpretabilidade, onde o papel de cada *feature* na decisão pode ser inspecionado, permite avaliar como uma abordagem linear simples pode capturar diferenças relevantes entre distintos tipos de texto.

A diversidade entre esses algoritmos permite melhor avaliar qual modelo é mais eficaz na detecção de textos gerados por IA, assim como observar as trocas entre complexidade, tempo de treinamento e interpretabilidade dos resultados. Detalhes técnicos como o pré-processamento dos dados, vetorização e configurações de treinamento foram adaptados para cada modelo.

2.3 PRÉ-PROCESSAMENTO

O pré-processamento dos dados textuais foi adaptado conforme os requisitos de cada arquitetura de modelo. Inicialmente, todos os textos foram convertidos para letras minúsculas, uma prática comum de padronização textual. Para os modelos XGBoost, Regressão Logística e SVM, essa etapa foi seguida pela vetorização dos textos com base na técnica *Term Frequency-Inverse Document Frequency* (TF-IDF). No caso do modelo XGBoost, utilizou-se um Pipeline composto por um *CountVectorizer* com max_features=5000 para limitar o vocabulário, seguido por um *TfidfTransformer* para ponderar as frequências dos termos. Já para Regressão Logística e SVM, aplicou-se diretamente o *TfidfVectorizer*, também com um limite de 5.000.

Para o modelo BERT, o pré-processamento exigiu etapas adicionais específicas à sua arquitetura. Foi implementada uma função de limpeza (*clean_text*) que removeu URLs, eliminou caracteres não alfanuméricos (mantendo os espaços) e converteu o texto para minúsculas. Em seguida, os textos foram tokenizados utilizando o *BertTokenizer* correspondente ao modelo pré-treinado *bert-base-multilingual-cased*, com truncamento até o comprimento máximo de 128 tokens (max_len=128).

2.4 DIVISÃO DOS CONJUNTOS DE DADOS

Foi construído um subconjunto balanceado, composto por 30.000 textos de cada classe, totalizando 60.000 instâncias, as quais foram embaralhadas de forma aleatória. Este mesmo conjunto foi utilizado para o treinamento e avaliação de todos os modelos considerados.

A divisão dos dados foi realizada com 80% para treinamento e 20% para teste ou validação, utilizando a função *train_test_split* da biblioteca *scikit-learn*, com *random_state=42* para garantir a reprodutibilidade dos experimentos. Para os modelos que permitem, como o BERT e o XGBoost, aplicou-se estratificação, de modo a preservar a proporção original entre as classes nos subconjuntos gerados.

No caso específico do modelo BERT, foi adotada a biblioteca PyTorch, e os dados foram organizados em lotes (batches) por meio da classe *DataLoader*, com o tamanho de lote (batch_size) configurado em 8 durante o treinamento.

2.5 VALIDAÇÃO

O desempenho dos modelos de detecção foi avaliado utilizando métricas padronizadas de classificação. Durante o treinamento, os modelos BERT, XGBoost, Regressão Logística e SVM foram avaliados sobre um conjunto de teste, obtido a partir da divisão estratificada do subconjunto balanceado de 60.000 textos.

Para todos os modelos, foi gerado um relatório de classificação proveniente da biblioteca *sklearn.metrics*, contemplando as métricas de precisão, revocação e F1-score por classe. A precisão mede a proporção de previsões corretas entre todas as previsões positivas realizadas. Já a revocação indica a proporção de casos positivos corretamente identificados pelo modelo. O F1-score representa a média harmônica entre precisão e revocação.

Além disso, foi calculada a matriz de confusão, a qual permitiu uma análise detalhada da natureza dos acertos e dos erros cometidos pelos modelos, indicando a quantidade de previsões corretamente classificadas como verdadeiros positivos e verdadeiros negativos, bem como os casos de classificações incorretas, representados por falsos positivos e falsos negativos.

3 DISCUSSÃO E RESULTADOS

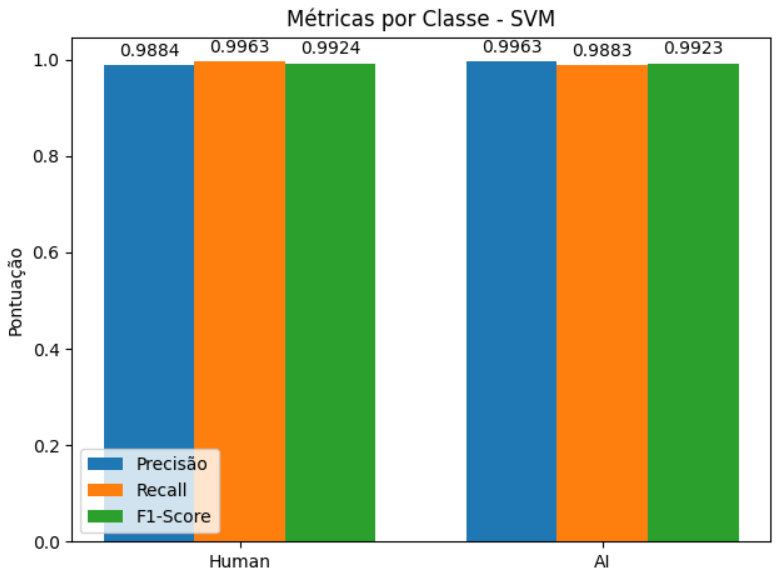
Este estudo comparou quatro abordagens diferentes para identificar textos gerados por inteligência artificial: SVM, Regressão Logística, XGBoost e BERT. Cada modelo foi avaliado usando um subconjunto balanceado de 60 mil textos, divididos igualmente entre textos humanos e textos gerados por IA.

Os resultados indicaram que os modelos tradicionais baseados em vetores TF-IDF apresentaram desempenho consistente, com destaque para o XGBoost e o SVM. O modelo BERT, embora mais sofisticado, demonstrou resultados competitivos, mesmo com uma configuração padrão de treinamento e pré-processamento simplificado. Isso sugere que, nesse cenário específico, tanto abordagens baseadas em representações vetoriais quanto métodos baseados em *embeddings* contextuais podem ser eficazes na detecção de textos gerados por IA.

3.1 SUPPORT VECTOR MACHINE

Como mostrado na Figura 7, o modelo SVM apresentou um desempenho altamente satisfatório, com métricas equilibradas e elevadas para ambas as classes. A acurácia final atingiu 99,23%, e os valores de precisão, recall e F1-score ultrapassaram 0,99 ou se mantiveram muito próximos disso, tanto para os textos escritos por humanos quanto para os gerados por IA.

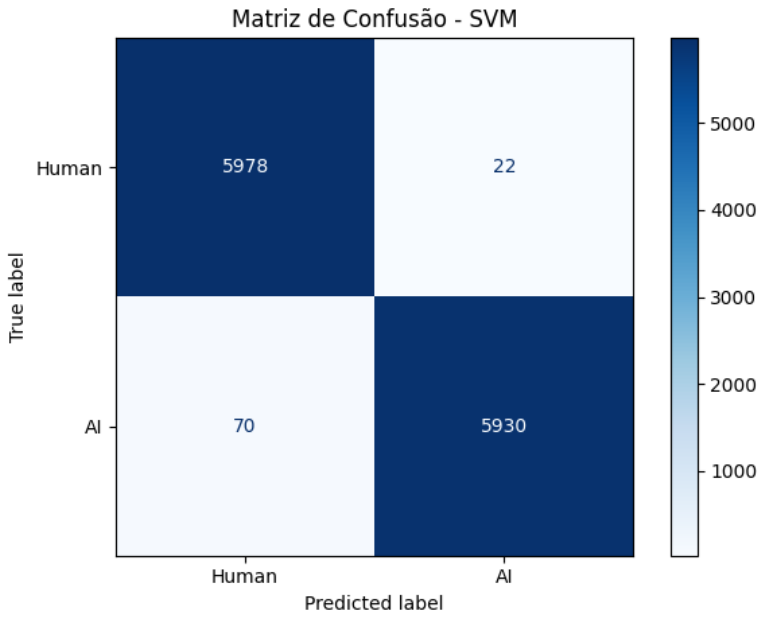
Figura 7 - Métricas do modelo SVM.



Fonte: Elaborado pelo autor.

Já a Figura 8 exibe a matriz de confusão, que demonstra a capacidade do modelo em distinguir corretamente as duas categorias, com um número reduzido de falsos positivos e falsos negativos.

Figura 8 - Matriz de confusão do modelo SVM.

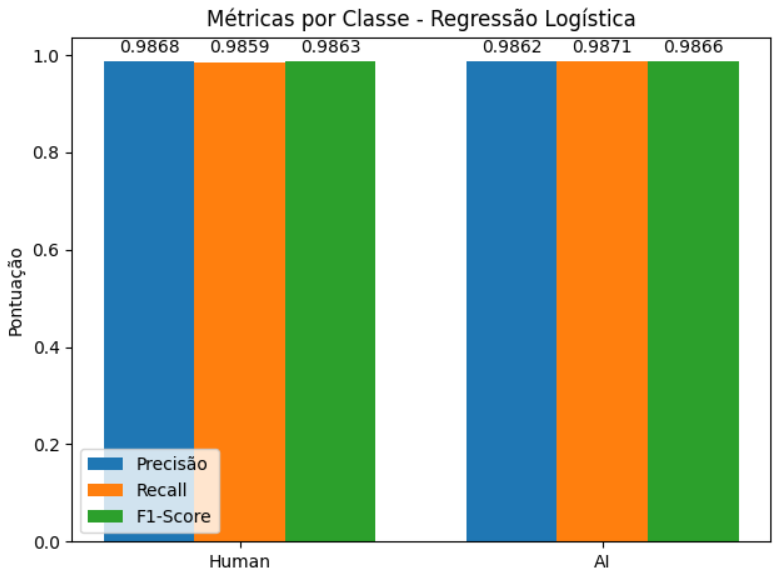


Fonte: Elaborado pelo autor.

3.2 REGRESSÃO LOGÍSTICA

A Figura 9 apresenta os resultados obtidos com o modelo de Regressão Logística, que demonstrou desempenho consistente na tarefa de classificação. A acurácia foi de 98,65%, com métricas de precisão, recall e F1-score bastante equilibradas entre as duas classes e levemente inferiores às obtidas pelo modelo SVM. Reforçando a Regressão Logística como uma ferramenta robusta, cujo desempenho em cenários de classificação pode se manter competitivo mesmo quando comparado a modelos de aprendizado de máquina considerados mais complexos.

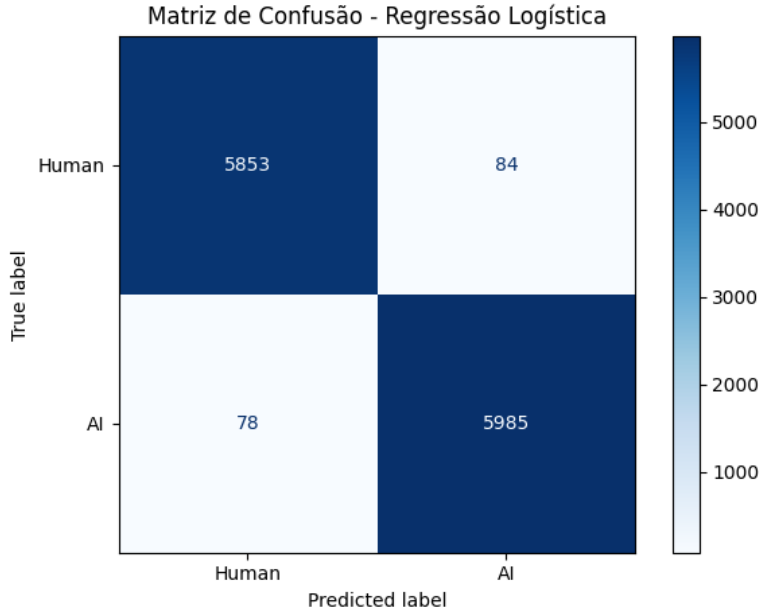
Figura 9 - Métricas do modelo Regressão Logística.



Fonte: Elaborado pelo autor.

A Figura 10 exibe a matriz de confusão, que evidencia um número moderado de erros de classificação, com distribuição relativamente homogênea entre falsos positivos e falsos negativos. Apesar de sua simplicidade, o modelo se mostrou competitivo e eficaz na distinção entre textos humanos e gerados por IA.

Figura 10 - Matriz de confusão do modelo Regressão Logística.

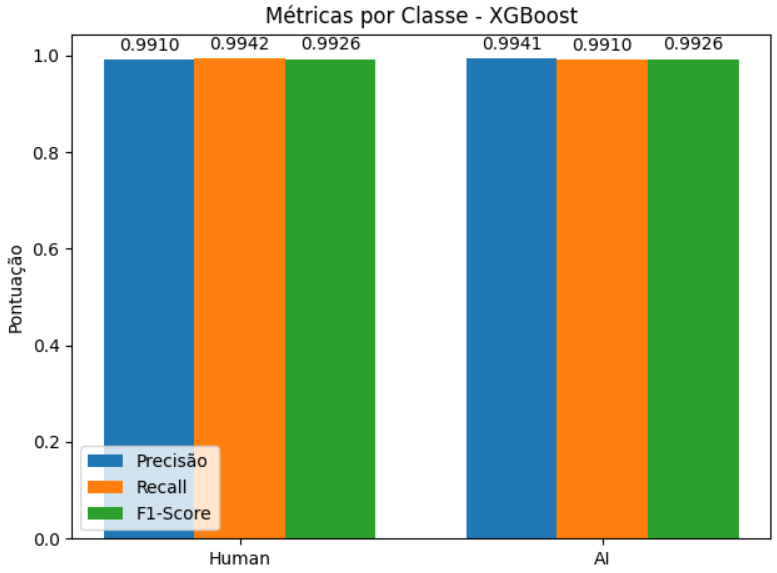


Fonte: Elaborado pelo autor.

3.3 XGBOOST

Conforme a Figura 11, o modelo XGBoost apresentou desempenho elevado, alcançando acurácia igual a 99,26%. As métricas de precisão, recall e F1-score foram equilibradas e consistentemente altas para ambas as classes.

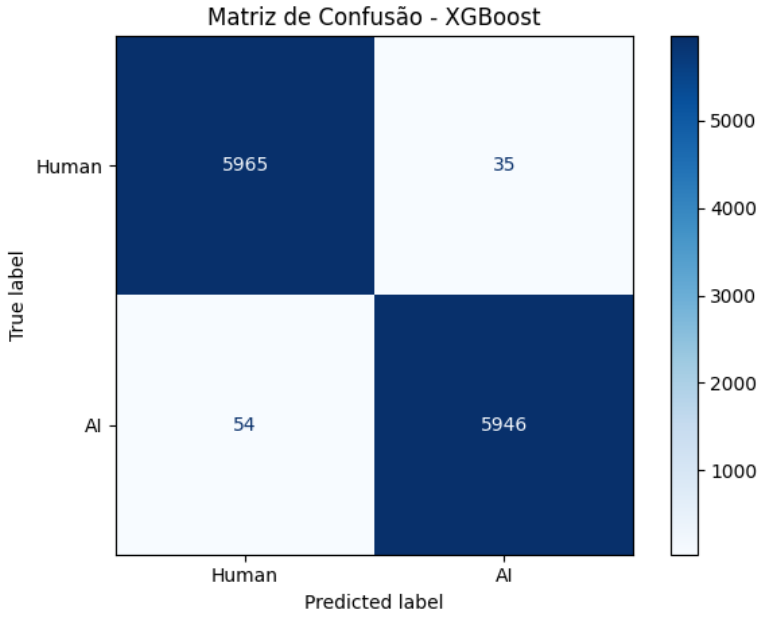
Figura 11 - Métricas do modelo XGBoost.



Fonte: Elaborado pelo autor.

A Figura 12, que apresenta a matriz de confusão, reforça a eficácia do modelo ao demonstrar sua capacidade de diferenciar com precisão entre textos humanos e gerados por IA, com baixa incidência de classificações incorretas.

Figura 12 - Matriz de confusão do modelo XGBoost.

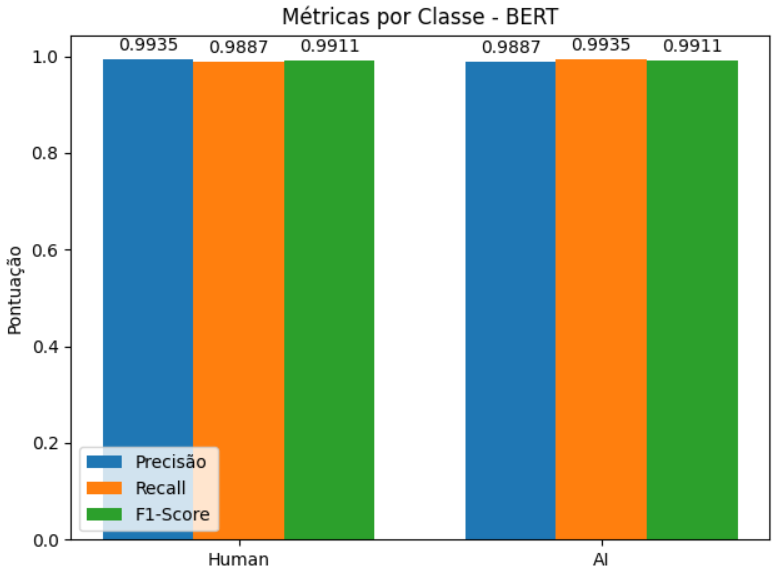


Fonte: Elaborado pelo autor.

3.4 BERT

A Figura 13 apresenta os resultados obtidos com o modelo BERT após duas épocas de treinamento com dados balanceados. Embora utilize uma arquitetura sofisticada baseada em aprendizado profundo, o modelo demonstrou desempenho semelhante aos demais métodos testados, atingindo acurácia de 99,11% e métricas de precisão, recall e F1-score bem equilibradas entre as duas classes. O número de épocas foi limitado a duas devido à observação de indícios iniciais de overfitting a partir da terceira época, momento em que a perda de validação começou a aumentar, comprometendo a generalização do modelo.

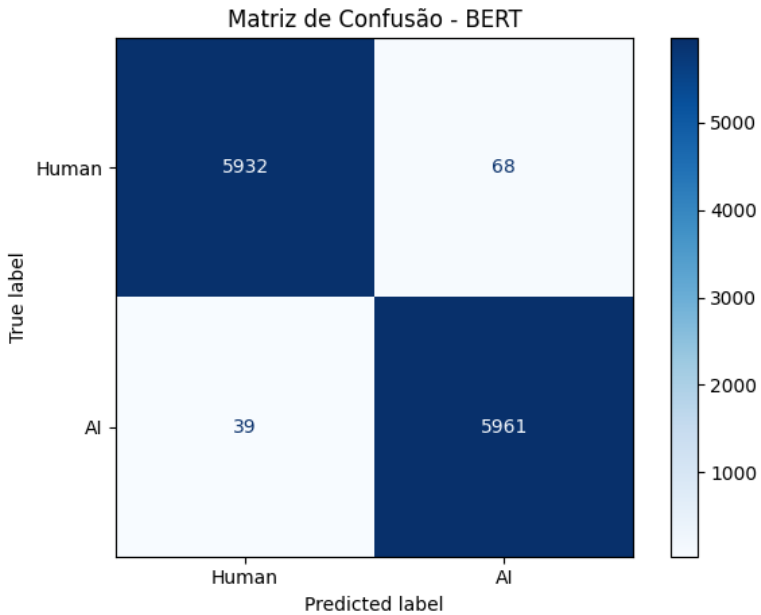
Figura 13 - Métricas do modelo BERT.



Fonte: Elaborado pelo autor.

A matriz de confusão da Figura 14 mostra que o modelo conseguiu distinguir com eficiência textos humanos de textos gerados por IA, apresentando baixo número de erros de classificação e resultados consistentes entre falsos positivos e falsos negativos.

Figura 14 - Matriz de confusão do modelo BERT.



Fonte: Elaborado pelo autor.

Comparando com estudos anteriores, os resultados reforçam as conclusões de Desaire et al. (2023) e Shah et al. (2023), que apontam o XGBoost como altamente eficiente na detecção de textos produzidos pelo ChatGPT, e também estão alinhados com os achados de Cingillioglu (2023), que destacam o potencial do SVM. Ainda assim, esta pesquisa apresenta limitações relevantes, como a ausência de validação cruzada, o uso de uma única base de dados e restrições computacionais que limitaram um ajuste mais aprofundado do modelo BERT. Apesar disso, os resultados obtidos oferecem uma base sólida para futuras aplicações práticas e abrem espaço para investigações com modelos híbridos ou treinamentos mais extensos.

4 CONCLUSÃO

Este trabalho teve como objetivo geral avaliar comparativamente métodos para a identificação de textos criados por modelos de inteligência artificial, analisando suas capacidades de classificação, desempenho e aplicabilidade. Para isso, foram selecionados e implementados quatro métodos distintos: BERT, XGBoost, SVM e Regressão Logística, aplicados sobre um subconjunto balanceado da base "AI Vs Human Text" (Gerami, 2024), composta por 60.000 textos rotulados.

O objetivo foi atingido com a construção de pipelines específicos para cada modelo, utilizando pré-processamento padronizado e avaliação com métricas consistentes. Todos os modelos apresentaram desempenho elevado, com acurácias acima de 98%, com XGBoost e SVM se destacando com mais de 99% de acurácia, enquanto o BERT, embora mais custoso computacionalmente, obteve resultados semelhantes.

Isso evidencia que modelos tradicionais baseados em TF-IDF ainda competem com arquiteturas sofisticadas, especialmente em bases bem estruturadas e rotuladas. Assim, a escolha do modelo ideal deve considerar, além da acurácia, fatores como custo computacional, interpretabilidade e eficiência.

Como continuidade deste trabalho, recomenda-se a expansão da análise para cenários mais complexos e realistas. Assim como a avaliação de textos híbridos, que mesclam conteúdo humano e de IA. Sugere-se também ampliar o escopo com métodos como o *Ghostbuster*, realizar validação cruzada mais rigorosa e testar textos modificados ou de modelos linguísticos mais recentes. Também é pertinente investigar técnicas de explicabilidade para compreender melhor os critérios de decisão adotados por cada classificador.

REFERÊNCIAS

ABDALI, S. et al. **Decoding the ai pen: Techniques and challenges in detecting ai-generated text**. 2024, ACM, p. 6428–6436.

ACHEAMPONG, F. A.; NUNOO-MENSAH, H.; CHEN, W. **Transformer models for text-based emotion detection: a review of bert-based approaches**. Artificial Intelligence Review, 2021, Springer, v. 54, n. 8, p. 5789–5829.

AKRAM, A. **An empirical study of ai generated text detection tools**. 2023, Opast Group LLC.

ALAMMAR, J. **The illustrated bert, elmo, and co**. How NLP cracked transfer learning, 2018, v. 38.

ALAMMAR, J.; GROOTENDORST, M. **Hands-on Large Language Models: Language Understanding and Generation**. [S.l.]: O'Reilly Media, Incorporated, 2024. ISBN 9781098150969.

ALMEIDA, J. C. P. de. **Textos gerados por inteligência artificial e suas implicações no ead**. EaD em Foco, 2023, v. 13, n. 1, p. e2083–e2083.

ARAUJO, M. **O uso de inteligência artificial para a geração automatizada de textos acadêmicos: plágio ou meta-autoria?** LOGEION: Filosofia da informação, 2016, v. 3, n. 1, p. 89–107.

BATISTELA, M. et al. **Inteligência artificial (ia): o desafio para driblar o plágio do homem e agora da máquina**. 2023, Galoá.

CHAUHAN, V. K.; DAHIYA, K.; SHARMA, A. **Problem formulations and solvers in linear svm: a review**. Artificial Intelligence Review, 2019, Springer, v. 52, n. 2, p. 803–855.

CINGILLIOGLU, I. **Detecting ai-generated essays: the chatgpt challenge**. The International Journal of Information and Learning Technology, 2023, Emerald Publishing Limited, v. 40, n. 3, p. 259–268.

DESAIRE, H. et al. **Accurately detecting ai text when chatgpt is told to write like a chemist**. Cell Reports Physical Science, 2023, Elsevier, v. 4, n. 11.

FERRO, V. d. S. et al. **As obras artísticas geradas pela inteligência artificial: considerações e controvérsias**. 2019, Universidade do Estado do Rio de Janeiro.

FEUERRIEGEL, S. et al. **Generative ai**. Business & Information Systems Engineering, 2024, Springer, v. 66, n. 1, p. 111–126.

FOLETTO, L. **Criação e cultura livre na era da inteligência artificial generativa**. Aurora. Revista de Arte, Mídia e Política, 2023, v. 16, n. 48, p. 76–92.

GAYAM, S. R. **Generative ai for content creation: Advanced techniques for automated text generation, image synthesis, and video production**. Journal of Science & Technology, 2022, v. 3, n. 1, p. 8–38.

GERAMI, S. **AI Vs Human Text**. 2024. Acesso em: 10 de out. 2024. Disponível em: <https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text>.

JIANG, H. et al. **Machine-learning-based user position prediction and behavior analysis for location services**. Information, 2021, MDPI, v. 12, n. 5, p. 180.

KAPRI, A. **Everything One Should Know About Support Vector Machines (SVM)**. 2020. Acesso em: 07 de mai. 2025. Disponível em: <https://medium.com/analytics-vidhya/everything-one-should-know-about-support-vector-machines-svm-18e6d3f96f49>.

KUMAR, R.; MINDZAK, M. **Who wrote this? detecting artificial intelligence–generated text from human-written text**. Canadian Perspectives on Academic Integrity, 2024, v. 7, n. 1.

MAHESH, B. **Machine learning algorithms-a review**. International Journal of Science and Research (IJSR).[Internet], 2020, v. 9, n. 1, p. 381–386.

MIYAZAKI, Y. et al. **Logistic regression analysis and machine learning for predicting post-stroke gait independence: a retrospective study**. Scientific reports, 2024, Nature Publishing Group UK London, v. 14, n. 1, p. 21273.

OTTEN, N. V. **Support Vector Machines (SVM) Made Simple & How To Tutorial**. 2024. Acesso em: 07 de maio 2025. Disponível em: <https://spotintelligence.com/2024/05/06/support-vector-machines-svm/>.

RAMRAJ, S. et al. **Experimenting xgboost algorithm for prediction and classification of different datasets**. International Journal of Control Theory and Applications, 2016, v. 9, n. 40, p. 651–662.

SANTOS, A. M. d. et al. **Usando redes neurais artificiais e regressão logística na predição da hepatite a**. Revista Brasileira de Epidemiologia, 2005, SciELO Public Health, v. 8, p. 117–126.

SARKAR, A. et al. **Text classification using support vector machine**. International Journal of Engineering Science Invention, 2015, v. 4, n. 11, p. 33–37.

SHAH, A. et al. **Detecting and unmasking ai-generated texts through explainable artificial intelligence using stylistic features**. International Journal of Advanced Computer Science and Applications, 2023, Science and Information (SAI) Organization Limited, v. 14, n. 10.

SHIJAKU, R.; CANHASI, E. **Chatgpt generated text detection**. 2023.

UZUN, L. **Chatgpt and academic integrity concerns: Detecting artificial intelligence generated content**. Language Education and Technology, 2023, v. 3, n. 1.

VERMA, V. et al. **Ghostbuster: Detecting text ghostwritten by large language models**. 2024, Association for Computational Linguistics, p. 1702–1717.

WANG, H.; LI, J.; LI, Z. **Ai-generated text detection and classification based on bert deep learning algorithm**. Theoretical and Natural Science, 2024, v. 39, p. 311–316.

WU, J. et al. **A survey on LLM-generated text detection: Necessity, methods, and future directions**. Computational Linguistics, 2025, MIT Press, Cambridge, MA, v. 51, n. 1.