XAI: ESCLARECENDO O PROBLEMA DA CAIXA PRETACOM TRANSPARÊNCIA E INTERPRETABILIDADE

XAI: CLARIFYING THE BLACK BOX PROBLEM WITH TRANSPARENCY AND INTERPRETABILITY

Gal Levy[1]
Mario Henrique Adaniya[2]

RESUMO

A crescente presença de tecnologias de Inteligência Artificial (IA) tem levantado questões acerca da transparência e compreensibilidade dos processos decisórios tomados por essas máquinas. O termo "caixa preta" é frequentemente utilizado para descrever sistemas de IA que operam de forma opaca, sem fornecer explicações claras sobre suas decisões. O objetivo desta pesquisa é analisar os métodos utilizados para resolver o problema da caixa preta e discutir as questões éticas e de preconceito que podem estar presentes nesses sistemas. Para tanto, foi realizada uma revisão bibliográfica dos principais métodos de interpretação de modelos de IA, tais como análise de sensibilidade, decomposição de modelo e interpretabilidade por design. A metodologia utilizada incluiu a seleção de artigos com base em critérios pré-definidos, extração de dados e análise dos resultados. Além disso, discutimos a importância da abordagem de equidade e justiça para garantir que os sistemas de IA não reproduzam preconceitos existentes na sociedade. Concluímos que os métodos de interpretabilidade podem ajudar a resolver o problema da caixa preta, mas devem ser combinados com abordagens de equidade e justiça para garantir a confiabilidade e responsabilidade dos sistemas de IA.

Palavras-chave: inteligência artificial; caixa preta; interpretabilidade; explicabilidade; preconceito.

ABSTRACT

The increasing presence of artificial intelligence (AI) technologies has raised questions about the transparency and understandability of the decision-making processes taken by these machines. The term "black box" is often used to describe AI systems that operate opaquely, without providing clear explanations about their decisions. The objective of this research is to analyze the methods used to solve the black box problem and discuss the ethical and prejudice issues that may be present in these systems. To do so, a bibliographic review of the main methods of interpreting AI models was carried out, such as sensitivity analysis, model decomposition, and interpretability by design. The methodology used included the

[1] Centro Universitário Filadélfia de Londrina - UniFil
[2] Centro Universitário Filadélfia de Londrina - UniFil

selection of articles based on pre-defined criteria, data extraction, and results analysis. Additionally, we discussed the importance of equity and justice approaches to ensure that AI systems do not reproduce existing biases in society. We concluded that interpretability methods can help solve the black box problem, but they should be combined with equity and justice approaches to ensure the reliability and accountability of AI systems.

Keywords: artificial intelligence; black box; interpretability; explainability; bias.

INTRODUÇÃO

A adoção da Inteligência Artificial (IA) tem levantado preocupações sobre a "caixa preta" associada aos sistemas de IA, especialmente aqueles baseados em redes neurais e aprendizado de máquina. Esses sistemas operam de forma opaca, tornando nebuloso como chegam às suas decisões. No entanto, pesquisadores estão explorando abordagens de Explicabilidade em Inteligência Artificial *(Explainable Artificial Intelligence - XAI)* para lidar com esse problema.

O XAI é uma área de estudo que busca tornar os sistemas de IA mais transparentes e explicáveis. Os pesquisadores estão desenvolvendo técnicas e métodos quepermitam compreender como os sistemas de IA tomam suas decisões, fornecendo explicações claras e compreensíveis aos usuários.

Essas abordagens podem envolver a interpretação dos modelos de IA, a geração de explicações pós-decisão e o projeto de modelos mais interpretáveis desde oinício. Ao utilizar o XAI, espera-se que seja possível compreender melhor os processos decisórios dos sistemas de IA, melhorar a confiança nas decisões tomadas por esses sistemas e evitar possíveis preconceitos ou discriminações ocultas. Embora a áreado XAI esteja em constante desenvolvimento, sua aplicação promissora pode ajudar a superar o desafio da "caixa preta" na IA, tornando os sistemas mais transparentes, compreensíveis e responsáveis.

Segundo Molnar, Casalicchio e Bischl (2018), a falta de transparência e inter-pretabilidade de modelos de aprendizado de máquina tem sido um problema constante. Além disso, Adadi e Berrada (2018) argumentam que a falta de transparência em modelos de IA podem levar a preconceitos e desigualdades

sociais. Entretanto, como será discutido no artigo, o problema da caixa preta é complexo e, mesmo sendo um problema na área por muito tempo, ainda não foi resolvido. Alguns dos métodos mais populares para tentar resolver o problema são apresentados em Koh; Liang (2017), que apresentam a técnica de *Influence Functions* para entender as previsões de um modelo de caixa preta; (Arrieta *et al*., 2019), que discutem o crescente interesse na área de XAI; enquanto (Guidotti *et al*., 2018) realizam uma revisão dos métodos. Neste artigo, serão discutidos esses métodos de solução para a caixa pretae a importância de resolver esse problema.

De acordo com Bender *et al*. (2021), a preocupação com o viés e a desigualdade social em sistemas de IA tem se intensificado à medida que modelos de linguagem cada vez maiores são desenvolvidos. Em seu artigo "*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big*", os autores argumentam que modelos de linguagem, como o GPT-3, podem perpetuar e ampliar o preconceito e a desigualdade, em vez de ajudar a resolvê-los. Eles destacam que esses modelos podem capturar e reproduzir os preconceitos existentes em grandes conjuntos de dados, e que o uso desses modelos sem uma compreensão completa de suas limitaçõespode levar a consequências prejudiciais e imprevisíveis. Portanto, é crucial considerar não apenas a transparência dos modelos de IA, mas também suas implicações sociaismais amplas, incluindo o viés e a desigualdade social.

A questão do viés e da desigualdade social em sistemas de IA é uma preocupação crescente em várias áreas, incluindo medicina e direito. Os autores Howard; Borenstein (2018) também observam a perpetuação de um viés nos sistemas de IA. Por exemplo, na medicina, os algoritmos de IA podem ser treinados em dados tendenciosos que afetam desproporcionalmente certos grupos, levando a diagnósticos imprecisos ou acesso desigual aos cuidados de saúde. Da mesma forma, no campo jurídico, os sistemas de IA podem refletir os preconceitos de seus criadores e perpetuar práticas discriminatórias. Além disso, conforme mencionado no artigo, ossistemas de IA podem ter noções preconcebidas sobre raça, gênero e outros fatores sociais, o que pode resultar em consequências não intencionais que reforçam ainda mais as desigualdades existentes. É essencial abordar essas

questões para garantir que a IA seja desenvolvida e utilizada de maneira responsável e ética.

As Funções de Influência permitem a compreensão da importância relativa de cada ponto de dados no modelo. Elas fornecem uma medida de quão sensível é uma predição em relação aos pontos de dados e podem ser usadas para identificar pontos de dados que são mais críticos para o modelo. Um exemplo de abordagem baseada em Funções de Influência é o artigo de (Koh; Liang, 2017), que apresenta uma técnica para entender como as previsões são afetadas por diferentes pontos de dados.

O pacote “iml” para a linguagem R é uma ferramenta que oferece uma variedade de métodos para interpretar modelos de aprendizado de máquina agnóstico, sendo possível aplicar as métricas em algoritmos diferentes. Ele inclui alguns métodos, como: de interpretação global, que visa gerar uma aproximação de uma árvore de decisão para uma interpretação mais fácil do resultado; O valor de Shapley é uma medida que distribui de forma justa a contribuição de cada característica na previsão de um modelo, considerando suas interações. É uma maneira de atribuir importância individual às características de forma equitativa. e as matrizes de correlação de recursos, bem como métodos de interpretação local, como o valor de contribuição individual e o perfil ICE. Um exemplo de uso do pacote R "iml" é o artigo de (Molnar;Casalicchio; Bischl, 2018).

XAI é uma área de pesquisa que se concentra na criação de modelos de IA que possam ser facilmente explicados e compreendidos. XAI inclui várias técnicas, como modelos de IA baseados em regras, modelos de IA transparentes e modelos de IA que fornecem explicações textuais ou visuais para suas decisões. Um exemplo de pesquisa em XAI é o artigo de (Arrieta *et al*., 2019), que propõe uma taxonomia para explicar a IA e discute oportunidades e desafios para o desenvolvimento de IA responsável. Este artigo revisa as abordagens mais recentes para resolver o problema da caixa preta na IA e discute como essas abordagens, incluindo Funções de Influência, pacote R "iml" e XAI, podem esclarecer dos modelos de IA e tornando a IA mais confiável e responsável.

Nesta pesquisa, destacamos uma abordagem prática para demonstrar como a Explicabilidade em XAI pode ser aplicada eficazmente. Utilizaremos uma ferramenta chamada LIME (*Local Interpretable Model-agnostic Explanations*), uma técnica inovadora de XAI. No contexto deste estudo, LIME será empregado como uma ferramentachave para ilustrar como os modelos de IA podem ser interpretados e explicados de maneira local e compreensível. Ao utilizar LIME, pretendemos também mostrar comoessas decisões podem ser traduzidas em explicações claras e acessíveis para os usuários finais. Este enfoque prático não apenas enfatiza a teoria por trás da XAI, mas também demonstra sua aplicação real e impacto tangível na compreensão dos sistemas de inteligência artificial.

REVISÃO DA LITERATURA

Guidotti *et al.* (2018) explica que a “caixa preta” em Inteligência Artificial (IA) refere-se ao fato de que, em muitos casos, os modelos de IA e os sistemas que os utilizam podem ser muito complexos e difíceis de entender. Por exemplo, as redes neurais profundas (do inglês, *Deep Neural Network* - DNNs) são frequentemente usadas em tarefas de processamento de imagens e de fala, mas suas camadas de processamento e pesos podem ser extremamente complexos, dificultando entender como o modelo está tomando decisões.

Bender *et al*. (2021) disse que a falta de transparência em modelos de IA é um problema significativo, pois pode levar a decisões injustas ou imprecisas, especialmente quando esses modelos são usados em áreas sensíveis, como saúde e justiça. Como resultado, houve uma crescente preocupação em desenvolver técnicas para tornar os modelos de IA mais transparentes e interpretáveis.

Uma abordagem comum para abordar a caixa preta em IA é o uso de métodos de interpretabilidade, como a análise de sensibilidade e a geração de mapas de ativação. Essas técnicas ajudam a identificar quais entradas e pesos são mais importantes para as saídas do modelo, permitindo que os usuários compreendam como o modeloestá chegando a suas decisões (Arrieta *et al*., 2019).

A análise de sensibilidade é uma técnica para a leitura de dados em IA que

envolve a avaliação do impacto de cada entrada na saída do modelo, permitindo que os usuários entendam quais recursos ou variáveis são mais importantes para as de- cisões do modelo. Essa técnica é especialmente útil em modelos complexos, como redes neurais profundas (Arrieta *et al*., 2019).

Geração do mapa de ativação é uma técnica que ajuda a visualizar a saída das camadas intermediárias em um modelo, permitindo que os usuários entendam como o modelo está processando as informações e tomando decisões. Por exemplo, em uma tarefa de reconhecimento de imagem, a geração do mapa de ativação pode mostrar quais partes da imagem foram mais importantes para a decisão final do modelo (Ghodrati *et al*., 2015).

LIME emerge como uma poderosa ferramenta de XAI. LIME é uma abordagem modelo-agnóstico que se destaca ao oferecer explicações locais para previsões de modelos de aprendizado de máquina complexos. A metodologia LIME trabalha gerando interpretações facilmente compreensíveis para as previsões do modelo em questão, destacando quais características dos dados de entrada influenciaram decisões específicas. Ao criar modelos interpretativos locais para previsões individuais, LIME permite aos pesquisadores e praticantes entenderem não apenas o que um modelo previu, mas também por que ele fez essa previsão específica. Sua aplicação tem sido amplamente reconhecida em diversas áreas, desde diagnósticos médicos até análise de sentimentos em texto, contribuindo significativamente para a disseminação da XAI, na prática. O uso efetivo de LIME ilustra como essa técnica pode superar a complexidade dos modelos de IA, tornando-os mais transparentes e, portanto, mais confiáveis para os usuários finais.

Outra abordagem é o uso de modelos interpretativos, projetados para serem mais transparentes e explicáveis. Por exemplo, os modelos baseados em árvores de decisão ou regressão linear são frequentemente usados em tarefas de classificação, pois são relativamente fáceis de entender e explicar.

Além disso, foram desenvolvidos métodos de verificação formal para garantir a transparência e a justiça dos modelos de IA. Esses métodos usam técnicas formais de verificação para analisar as propriedades dos modelos de IA,

como a segurança e a privacidade, ajudando a identificar possíveis problemas antes que eles ocorram.

Essas técnicas são valiosas para entender como os modelos de IA estão tomando decisões e identificar possíveis problemas, como viés ou falta de interpretabilidade. No entanto, é importante lembrar que elas são apenas uma parte do processo de tornar os modelos de IA mais transparentes e interpretáveis, e que outras técnicas, como modelos interpretativos e verificação formal, também são importantes para garantir a imparcialidade e precisão dos modelos de IA em todas as situações.

Em resumo, embora a caixa preta em IA seja um problema desafiador, existem várias abordagens promissoras para tornar os modelos de IA mais transparentes e interpretáveis. No entanto, ainda há muito trabalho a ser feito para desenvolver técnicas eficazes e garantir que os modelos de IA sejam justos e precisos em todas as situações em que são utilizados.

METODOLOGIA

A metodologia adotada no trabalho pode ser resumida nos seguintes passos:

• **Identificação da Problemática**: Iniciamos nossa pesquisa com uma investigação ampla utilizando a plataforma baseada em IA, "Semanticscholar". Termos- chave como "IA", "caixa preta", "redes neurais" e "desigualdade em IA" foram empregados para identificar um conjunto inicial de artigos relevantes.

• **Seleção e Análise dos Artigos Iniciais**: O artigo (Arrieta *et al*., 2019) emergiu como peça central, servindo como a base conceitual do nosso trabalho. A partir dele, exploramos métodos e conceitos relacionados ao XAI para resolver o problema da "caixa preta". Uma análise crítica desses artigos iniciais permitiu-nos compreender as oportunidades e desafios associados ao campo.

• **Pesquisa Adicional e Exploração de Métodos XAI**: Continuamos nossa pesquisa procurando métodos específicos dentro do XAI que abordam a

opacidade dos modelos de IA. Aprofundamos nosso entendimento, explorando diferentes técnicas que foram empregadas para resolver a problemática da "caixa preta". Esta etapa foi crucial para identificar uma variedade de abordagens utilizadas na comunidade científica para tornar os modelos de IA mais interpretáveis e transparentes.

• **Experimentação e Aplicação Prática**: Para validar a eficácia dos métodos XAI identificados, realizamos um experimento prático. Utilizamos a técnica LIME sobre duas redes neurais, aplicando-a em avaliações para determinar palavras- chave indicativas de sentimentos positivos ou negativos. Esta experimentação forneceu visões tangíveis sobre como o XAI pode ser implementado na prática, evidenciando seu papel na interpretabilidade dos modelos de IA.

• **Análise e Síntese dos Resultados**: Após a experimentação, procedemos com uma análise detalhada dos resultados obtidos. Comparamos as palavras indica- tivas identificadas pelo LIME com as expectativas teóricas, avaliando a eficácia da abordagem. Esta análise nos permitiu fazer conclusões informadas sobre a utilidade e os desafios do XAI na resolução da "caixa preta" em sistemas de IA.

• **Elaboração do Trabalho**: Finalmente, baseados na revisão bibliográfica e nos resultados do experimento, desenvolvemos o presente trabalho. Nossa abordagem metodológica permitiu-nos explorar e contextualizar o estado atual do XAI, apresentando um panorama abrangente dos esforços para tornar a IA mais transparente.

DESENVOLVIMENTO

Nesta seção, exploramos os resultados de nossa análise, utilizando duas arquiteturas distintas de redes neurais para classificação de sentimentos em avaliações de filmes. A primeira arquitetura envolve uma rede neural simples com uma única camada oculta, enquanto a segunda é uma versão mais complexa com

duas camadas ocultas. Ambas as redes foram treinadas usando o conjunto de dados IMDB para diferenciar entre "avaliações positivas" e "avaliações negativas".

Durante o desenvolvimento deste estudo, empregamos metodologias robustas para otimizar a precisão e confiabilidade dos resultados. Utilizamos a função de perda BCELoss (*Binary Cross-Entropy Loss*), uma escolha comum em tarefas de classificação binária, como a nossa. A BCELoss é fundamental, pois calcula a discrepância entre as previsões do modelo e os rótulos reais, orientando o treinamento de forma eficaz. Para aprimorar ainda mais o desempenho, aplicamos o otimizador Adam. O otimizador Adam é especialmente poderoso, pois ajusta as taxas de aprendizado para cada parâmetro da rede, acelerando significativamente a convergência do modelo durante o treinamento. Essas escolhas auxiliam na garantia de uma abordagem sólida e confiável para nossas análises.

Tabela 1 – Principais palavras em análises positivas e negativas para a rede neural simples (1000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	15.69%
veri	12.95%
great	10.39%
also	9.76%
love	9.72%
film	9.59%
well	8.55%
see	8.00%
show	7.83%
stori	7.51%

Análises Negativas

Palavra-chave	Influência
thi	20.03%
wa	14.45%
movi	13.38%
even	10.49%
onli	7.84%
look	7.60%
would	7.44%
dont	7.13%
bad	6.09%
act	5.54%

Tabela 2 – Principais palavras em análises positivas e negativas para a rede neuralcomplexa (1000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	15.70%
veri	12.95%
great	10.37%
also	9.75%
love	9.72%
film	9.60%
well	8.56%
see	8.00%
show	7.86%
stori	7.50%

Análises Negativas

Palavra-chave	Influência
thi	20.07%
wa	14.45%
movi	13.38%
even	10.46%
onli	7.84%
look	7.58%
would	7.45%
dont	7.12%
bad	6.10%
act	5.55%

Em tarefas de processamento de linguagem natural, como análise de sentimento, os dados de texto passam por etapas de pré-processamento, onde palavras são transformadas em tokens, as unidades fundamentais usadas para análise. Durante a tokenização, palavras comuns são frequentemente abreviadas ou representa-das em suas formas reduzidas, ou lematizadas, como ’hi’ em vez de ’high’ e ’thi’ em vez de ’this’. Essa prática serve a diversos propósitos, incluindo a redução da complexidade computacional, o tratamento de variações de palavras e o aprimoramento da precisão das tarefas de análise de texto. Essas formas abreviadas mantêm o significado central das palavras, simplificando a análise e melhorando a eficiência e eficácia dos modelos de processamento de linguagem natural.

Tabela 3 – Principais palavras em análises positivas e negativas para a rede neural simples (2000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	19.40%
veri	12.41%
great	10.86%
well	10.84%
show	10.50%
stori	8.21%
love	7.82%
also	7.46%
see	6.48%
year	6.03%

Análises Negativas

Palavra-chave	Influência
wa	17.38%
thi	17.19%
movi	13.88%
even	10.62%
onli	8.38%
ani	6.86%
would	6.69%
like	6.60%
bad	6.48%
act	5.92%

Tabela 4 – Principais palavras em análises positivas e negativas para a rede neuralcomplexa (2000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	19.39%
veri	12.44%
great	10.84%
well	10.83%
show	10.50%
stori	8.21%
love	7.82%
also	7.47%
see	6.48%
year	6.02%

Análises Negativas

Palavra-chave	Influência
wa	17.38%
thi	17.20%
movi	13.88%
even	10.61%
onli	8.38%
ani	6.86%
would	6.69%
like	6.59%
bad	6.48%
act	5.92%

Ao examinarmos as palavras-chave essenciais que influenciam significativamente as decisões de classificação de cada modelo, uma revelação intrigante se destaca: as palavras-chave mais significativas para ambas as redes são surpreendentemente semelhantes. Esse fenômeno destaca a eficácia do método LIME em fornecer explicações precisas e coerentes, independentemente da complexidade da arquitetura subjacente.

Tabela 5 – Principais palavras em análises positivas e negativas para a rede neural simples (3000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	15.86%
veri	14.58%
film	11.38%
love	9.09%
well	9.05%
stori	9.02%
great	8.75%
also	8.66%
show	7.69%
play	5.91%

Análises Negativas

Palavra-chave	Influência
thi	16.95%
wa	16.57%
movi	15.10%
onli	8.99%
bad	8.68%
act	7.25%
like	6.83%
would	6.76%
even	6.53%
look	6.32%

Tabela 6 – Principais palavras em análises positivas e negativas para a rede neuralcomplexa (3000 amostras)

Análises Positivas

Palavra-chave	Influência
hi	15.86%
veri	14.58%
film	11.38%
love	9.09%
well	9.05%
stori	9.02%
great	8.75%
also	8.66%
show	7.69%
play	5.91%

Análises Negativas

Palavra-chave	Influência
thi	16.95%
wa	16.57%
movi	15.10%
onli	8.99%
bad	8.68%
act	7.25%
like	6.83%
would	6.76%
even	6.53%
look	6.32%

Essas novas percepções aprimoram nossa compreensão do cenário de interpretabilidade em aprendizado de máquina. A interpretabilidade tornou-se fundamental, especialmente em aplicações onde as decisões dos modelos impactam significativamente os usuários finais. Em contextos como diagnósticos médicos ou tomada de decisões financeiras, compreender os motivos por trás das previsões dos modelos é fundamental para estabelecer confiança e aceitação das decisões geradas pelo sistema.

Tabela 7 – Principais palavras em análises positivas e negativas para a rede neuralsimples (4000 amostras)

Análises Positivas

Palavra-chave	Influência
film	21.72%
hi	13.29%
veri	11.31%
great	9.68%
love	8.41%
stori	8.01%
also	7.61%
well	6.97%
see	6.59%
one	6.41%

Análises Negativas

Palavra-chave	Influência
thi	19.89%
wa	14.65%
even	11.52%
would	8.71%
look	8.19%
movi	8.18%
like	7.40%
bad	7.30%
onli	7.21%
dont	6.96%

Tabela 8 – Principais palavras em análises positivas e negativas para a rede neuralcomplexa (4000 amostras)

Análises Positivas

Palavra-chave	Influência
film	21.72%
hi	13.28%
veri	11.32%
great	9.68%
love	8.40%
stori	8.01%
also	7.60%
well	6.96%
see	6.61%
one	6.40%

Análises Negativas

Palavra-chave	Influência
thi	19.89%
wa	14.65%
even	11.53%
would	8.70%
look	8.20%
movi	8.17%
like	7.39%
bad	7.30%
onli	7.22%
dont	6.95%

Neste cenário mais amplo, o LIME se destaca como uma ferramenta inestimável. Sua habilidade em identificar palavras específicas em avaliações de filmes que impactam a polaridade do sentimento proporciona visões tangíveis. Notavelmente, a palavra "film" emerge como um dos indicadores mais significativos para avaliação positiva, indicando que referências específicas à palavra "filme" em uma avaliação podem ser marcadores cruciais de sentimento. Por outro lado, a

palavra "movie" aparece repetidamente como um indicador de avaliação negativa, destacando seu papel como um termo associado a sentimentos menos favoráveis nas análises de filmes. Essa distinção revela nuances importantes na forma como as palavras são percebidas e comoinfluenciam a avaliação global dos filmes.

Tabela 9 – Principais palavras em análises positivas e negativas para a rede neuralsimples (5000 amostras)

Análises Positivas

Palavra-chave	Influência
veri	14.06%
hi	13.87%
film	11.14%
great	10.97%
see	9.57%
love	8.97%
also	8.19%
stori	7.86%
well	7.77%
show	7.59%

Análises Negativas

Palavra-chave	Influência
movi	20.49%
wa	17.72%
thi	13.11%
like	9.67%
even	8.59%
onli	6.93%
look	6.27%
would	6.03%
make	6.01%
seem	5.19%

Tabela 10 – Principais palavras em análises positivas e negativas para a rede neuralcomplexa (5000 amostras)

Análises Positivas

Palavra-chave	Influência
film	14.68%
veri	13.63%
hi	13.26%
great	10.47%
show	9.02%
love	8.37%
well	8.11%
also	7.82%
see	7.43%
ha	7.22%

Análises Negativas

Palavra-chave	Influência
thi	19.17%
wa	14.97%
movi	14.24%
would	8.44%
like	8.04%
onli	7.92%
act	7.73%
even	7.54%
bad	6.18%
ani	5.77%

Essas descobertas destacam a eficiência do LIME em fornecer explicações precisas, independentemente da complexidade do modelo subjacente. A capacidade de destacar palavras-chave cruciais para a classificação de sentimentos

é fundamental para a compreensão e interpretação das decisões dos modelos de aprendizado de máquina, oferecendo visões inestimáveis em aplicações do mundo real.

CONCLUSÃO

Neste estudo, investigamos o ponto de encontro intrigante entre Inteligência Artifical (IA) e *Explanable Artificial Intelligence* (XAI), uma área de pesquisa cada vez mais crucial. Ao examinar a complexa relação entre as "caixas pretas" dos modelos de IA e a crescente necessidade de entender e explicar seus processos internos, mergulhamos profundamente na essência da XAI. Esta exploração auxilia no esclarecimento do cenário atual, e também trouxe uma compreensão essencial sobre a urgência de transparência e compreensibilidade em sistemas automatizados.

Ao aplicarmos a técnica de LIME, em um contexto prático, conseguimos desvendar algumas das complexidades das "caixas pretas" da IA. Demonstramos empiricamente como a interpretabilidade não é apenas um campo de estudo isolado, mas sim um pilar fundamental no desenvolvimento contínuo de modelos de IA. A XAI não apenas nos ajuda a entender as decisões dos modelos, mas também abre portas para a melhoria constante desses modelos no futuro.

A importância contínua de estudar a interação entre IA e XAI não pode ser subestimada. À medida que continuamos a avançar nas fronteiras do conhecimento, é imperativo dedicar recursos e esforços para aprimorar nossa compreensão desses sistemas complexos. Somente através da pesquisa contínua e do aprimoramento de técnicas como LIME, poderemos enfrentar os desafios cada vez maiores da IA, garantindo que ela não apenas beneficie a humanidade, mas também seja acessível, compreensível e ética. Ao investirmos na pesquisa em XAI, estamos moldando o futuro da inteligência artificial, tornando-o mais transparente, responsável e preparado para enfrentar as demandas do mundo moderno.

À medida que nos aprofundamos na interseção complexa entre Inteligência Artificial e Explicabilidade, uma miríade de oportunidades para futuras

pesquisas emerge. Uma área promissora é a exploração de técnicas XAI mais avançadas, como SHAP (*Shapley Additive Explanations*) e LRP (*Layer-wise Relevance Propagation*), para aprofundar nossa compreensão dos modelos de IA. Além disso, investigar como a interpretabilidade pode ser integrada no ciclo de vida do desenvolvimento de IA, desde a concepção até a implementação prática, pode fornecer visões valiosos. Além disso, considerando a rápida evolução dos modelos de IA, é imperativo examinar como as técnicas de XAI podem ser adaptadas para lidar com arquiteturas emergentes, como redes neurais profundas e modelos de aprendizado federado. Ao abordar esses desafios, podemos continuar avançando em direção a sistemas de IA que são não apenas poderosos, mas também compreensíveis e éticos, garantindo um futuro mais transparente e confiável.

REFERÊNCIAS

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018.

ARRIETA, A. B. *et al*. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *ArXiv*, abs/1910.10045, 2019.

BENDER, E. M. *et al*. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

GHODRATI, A. *et al.* Deepproposal: Hunting objects by cascading deep convolutional layers. *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 2578–2586, 2015.

GUIDOTTI, R. *et al*. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, v. 51, p. 1 – 42, 2018.

HOWARD, A. M.; BORENSTEIN, J. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, v. 24, p. 1521–1536, 2018.

KOH, P. W.; LIANG, P. Understanding black-box predictions via influence functions. *ArXiv*, abs/1703.04730, 2017.

MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. iml: An r package for interpretable machine learning. *J. Open Source Softw.*, v. 3, p. 786, 2018.