IMAT: Uma Ferramenta para Análise de Modelos de Aprendizado de Máquina Interpretáveis

André Assis[1], **Jamilson Dantas** [2], **Ermeson Andrade**[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Recife, PE – Brasil

{andre.assis, ermeson.andrade}@ufrpe.br

[2]Centro de Informática – Universidade Federal de Pernambuco
Recife, PE – Brasil

jrd@cin.ufpe.br

Abstract. *Transparency and interpretability of Artificial Intelligence (AI) and Machine Learning (ML) models are increasingly relevant in applications involving social network analysis and data mining. While advanced models, such as deep learning, provide robust solutions to complex problems, their growing complexity makes it difficult to understand the decisions they make. The lack of transparency can undermine user trust and hinder the widespread adoption of these technologies. To address this challenge, this paper presents IMAT (Interpretable Models Analysis Tool), a tool designed to generate flowcharts that map each stage of data processing in deep learning models. IMAT aims to provide a clear and accessible visualization of the data flow and internal operations of models, from data input to output generation, facilitating their interpretation. Additionally, this work discusses the architecture and functionalities of IMAT and demonstrates its application in sentiment analysis of tweets, using the MLP (MultiLayer Perceptron) algorithm, evaluating the implications and limitations of the obtained results.*

Resumo. *A transparência e interpretabilidade dos modelos de Inteligência Artificial (IA) e Machine Learning (ML) são cada vez mais relevantes em aplicações que envolvem análise de redes sociais e mineração de dados. Embora modelos avançados, como os de deep learning, ofereçam soluções robustas para problemas complexos, sua crescente complexidade dificulta a compreensão das decisões tomadas. A falta de transparência pode comprometer a confiança dos usuários e limitar a adoção dessas tecnologias. Para enfrentar esse desafio, este artigo apresenta a IMAT (Interpretable Models Analysis Tool), uma ferramenta desenvolvida para gerar fluxogramas que mapeiam cada etapa do processamento de dados em modelos de deep learning. A IMAT visa oferecer uma visualização clara e acessível do fluxo de dados e das operações internas dos modelos, desde a entrada até a geração da resposta, facilitando sua interpretação. Além disso, este trabalho discute a arquitetura e funcionalidades da IMAT e demonstra sua aplicação na análise de sentimentos em tweets, utilizando o algoritmo MLP (MultiLayer Perceptron), avaliando as implicações e limitações dos resultados obtidos.*

1. Introdução

A Inteligência Artificial (IA) e a aprendizagem de máquina têm se tornado fundamentais em diversas indústrias e tecnologia da informação em geral. Visando solucionar problemas complexos do mundo real, os modelos robustos de *deep learning* destacam-se por sua capacidade de realizar previsões precisas a partir de grandes volumes de dados. No entanto, a crescente complexidade desses modelos levanta questões críticas sobre a interpretabilidade e transparência das decisões tomadas, aspectos que são essenciais para a aceitação e confiabilidade da tecnologia, especialmente em áreas sensíveis como diagnósticos médicos e decisões financeiras. A falta de transparência nos modelos pode resultar em desconfiança por parte dos usuários finais, comprometendo a adoção generalizada dessas tecnologias. Além disso, a interpretabilidade é importante para identificar e corrigir vieses e erros nos modelos, garantindo a justiça e a precisão das previsões [Assis et al. 2023]. Essa necessidade de explicações que facilitem a compreensão do modelo se torna ainda mais evidente em aplicações que lidam com dados não estruturados e altamente subjetivos, como os encontrados em redes sociais.

As redes sociais, especialmente o Twitter (atualmente denominado X), tornaram-se ambientes ricos para a análise de dados textuais, possibilitando a exploração de opiniões públicas sobre uma ampla variedade de temas [Silva et al. 2021, Paes et al. 2022]. A análise de sentimentos com modelos de IA, que identifica e classifica emoções expressas em textos, tem se mostrado particularmente valiosa para compreender percepções sociais em tempo real. Aplicável a grandes volumes de *tweets*, essa abordagem fornece informações relevantes para a tomada de decisões em áreas como marketing, política e saúde pública. No entanto, a correta interpretação dos resultados desses modelos exige atenção à transparência e confiabilidade. Compreender os critérios que levaram a uma determinada classificação é fundamental para validar os resultados, detectar vieses e garantir a qualidade das inferências produzidas.

Um dos principais desafios enfrentados é a dificuldade em compreender e explicar os processos internos de modelos de *deep learning*, que frequentemente operam como "caixas-pretas", tornando os caminhos que levam a determinadas previsões difíceis de entender, até mesmo para especialistas [Alves and ANDRADE 2022]. Essa falta de transparência impede a validação e a confiança nas previsões, além de dificultar a identificação de vieses e erros potenciais. Portanto, é fundamental desenvolver ferramentas que proporcionem uma compreensão clara e acessível do processo decisório desses modelos. Tais ferramentas interpretativas desempenham um papel crucial na validação e melhoria contínua dos modelos de ML, fornecendo *feedback* essencial para ajustes e refinamentos [Cortiz 2021]. Ao oferecer uma visão detalhada dos mecanismos internos dos modelos, elas possibilitam que desenvolvedores identifiquem padrões inadequados, corrijam vieses e realizem ajustes precisos [Agarwal and Das 2020]. Além disso, a integração de ferramentas de interpretabilidade em ciclos regulares de revisão e atualização contribui para a criação de sistemas mais justos, confiáveis e alinhados às expectativas e necessidades dos usuários.

Diversas abordagens têm sido propostas para enfrentar os desafios da interpretabilidade em modelos de aprendizado de máquina, com destaque para técnicas como LIME (*Local Interpretable Model-agnostic Explanations*), SHAP (*SHapley Additive exPlanations*) e DALEX (*Descriptive mAchine Learning EXplanations*). Essas técnicas ofer-

ecem explicações locais que ajudam a entender o comportamento dos modelos em instâncias específicas, tornando visíveis seus mecanismos internos e contribuindo para maior transparência em diferentes contextos de aplicação [Salih et al. 2023]. Apesar desses avanços, ainda existe um descompasso entre essas soluções e as necessidades de usuários finais não técnicos, como analistas de negócio e profissionais de domínio, que frequentemente encontram dificuldade na execução de *scripts* ou na interpretação de métricas estatísticas. Para esse público, explicações em forma de tabelas de importância ou gráficos abstratos não são suficientes, tornando essencial uma visualização clara que comunique o raciocínio do modelo de forma imediata e acessível. Além disso, persiste uma lacuna significativa na integração de ferramentas que, além de fornecerem explicações, visualizem todo o processo de decisão de maneira compreensível para os usuários finais, facilitando a interpretação em ambientes industriais e acadêmicos [Bhattacharya 2022].

Para preencher essa lacuna, este trabalho propõe o IMAT *(Interpretable Models Analysis Tool)*, uma ferramenta desenvolvida para construir fluxogramas detalhados e resumidos que mapeiam cada etapa do processamento de dados em modelos de *deep learning*. O IMAT busca oferecer uma visualização clara e acessível do fluxo de dados e das operações internas dos modelos, desde a entrada dos dados até a geração da resposta, facilitando a compreensão e aumentando a transparência dos modelos. Ao fornecer uma representação gráfica e detalhada do funcionamento interno dos modelos, o IMAT visa não apenas aumentar a confiança dos usuários, mas também fornecer uma ferramenta poderosa para a análise e melhoria contínua dos modelos de ML, atendendo às necessidades de diferentes usuários. Além disso, este trabalho tem como objetivos avaliar a aplicabilidade do IMAT por meio de um estudo de caso, no qual analisamos o funcionamento de um modelo de rede neural do tipo MLP aplicado à análise de sentimentos de *tweets*, utilizando o dataset *Sentiment140*.

A estrutura do artigo é organizada da seguinte forma: a Seção 2 introduz o conceito de Inteligência Artificial Explicável, destacando sua importância e técnicas principais. A Seção 3 revisa os principais trabalhos relacionados à interpretabilidade em ML, justificando a necessidade da IMAT. A Seção 4 descreve as principais funcionalidades da IMAT. A Seção 5 apresenta um estudo de caso aplicado à análise de sentimentos de *tweets*. Por fim, a Seção 6 apresenta considerações finais e sugestões para pesquisas futuras.

2. Inteligência Artificial Explicável

A XAI, do inglês *eXplainable Artificial Intelligence* é um campo de pesquisa que busca desenvolver métodos e técnicas que permitam a compreensão e interpretação dos modelos de inteligência artificial por seres humanos. O seu principal objetivo é reduzir a opacidade dos modelos de IA, frequentemente referidos como "caixas-pretas", e torná-los mais transparentes e confiáveis para os usuários finais [Adadi and Berrada 2018]. A opacidade dos modelos de IA impede que os humanos entendam os processos decisórios subjacentes, o que pode levar à desconfiança e à relutância em adotar essas tecnologias [Meske and Bunde 2020]. A transparência, por outro lado, refere-se à capacidade de um sistema de IA de ser compreendido por seus usuários, enquanto a explicabilidade envolve a capacidade do sistema de fornecer justificativas compreensíveis para suas decisões e previsões [Gregor and Benbasat 1999].

A interpretabilidade e a explicabilidade em Aprendizado de Máquina, segundo autores como Miller, Kim e Molnar, referem-se à capacidade de um humano entender e prever consistentemente as decisões de um modelo, tornando-o transparente e compreensível [Miller 2019, Kim et al. 2016, Molnar 2020]. Essa característica é fundamental para aumentar a confiança dos usuários em sistemas de IA, especialmente em domínios críticos como saúde, onde decisões algorítmicas podem ter consequências significativas. A XAI permite identificar e corrigir vieses, melhorar o desempenho dos modelos e garantir que eles operem de maneira justa e ética, contribuindo para a construção de sistemas de IA mais confiáveis e responsáveis [Sundar 2020].

Os modelos de aprendizado de máquina variam em sua complexidade e, consequentemente, em sua capacidade de explicação. Modelos mais simples são geralmente mais transparentes, permitindo uma compreensão mais intuitiva de seu funcionamento. Em contraste, modelos mais complexos tendem a ser opacos, dificultando a interpretação direta de seus processos internos [Angelov et al. 2021]. A pesquisa em XAI busca desenvolver métodos que aumentem a transparência e a interpretabilidade de ambos os tipos de modelos, fornecendo informações detalhadas sobre suas operações internas e decisões [Arrieta et al. 2020]. Promovendo assim, a confiança dos usuários e facilitando a adoção de sistemas de IA, ao mesmo tempo que assegura a responsabilidade e a transparência em aplicações críticas [Adadi and Berrada 2018].

No contexto das redes sociais e da mineração de dados, a XAI desempenha um papel crucial ao garantir transparência e confiabilidade nas análises. Plataformas como X geram grandes volumes de dados textuais, utilizados para identificar tendências e compreender opiniões públicas. A análise de sentimentos, por exemplo, é amplamente aplicada em marketing, política e saúde pública, mas sua eficácia depende da interpretabilidade dos modelos utilizados. A opacidade dos modelos de deep learning pode comprometer a aceitação de suas inferências, especialmente quando impactam políticas públicas ou estratégias empresariais [Søgaard 2023]. Métodos de XAI permitem compreender os critérios que levam a determinadas classificações, identificar vieses e garantir que os modelos operem de maneira justa e alinhada às expectativas dos usuários.

Além de melhorar a qualidade das análises, a integração de técnicas explicáveis na mineração de dados em redes sociais amplia a adoção dessas tecnologias. Ferramentas que oferecem visualizações claras do processo decisório dos modelos possibilitam que especialistas e não especialistas interpretem melhor os resultados e tomem decisões mais embasadas [Arrieta et al. 2020]. Assim, ao promover maior transparência e compreensibilidade, a XAI fortalece o uso ético e responsável da inteligência artificial em ambientes altamente dinâmicos.

3. Trabalhos Relacionados

A crescente complexidade dos modelos de aprendizado de máquina gerou uma demanda significativa por ferramentas que facilitem a compreensão de suas operações internas. Considerando que ferramentas de XAI são essenciais para garantir transparência, confiança e adoção ética da IA em diversos setores, esta seção revisa os principais trabalhos e ferramentas relacionadas à interpretabilidade em modelos de aprendizado de máquina, destacando as contribuições e as justificativas para o desenvolvimento da ferramenta IMAT.

Em [Angelov et al. 2021], é apresentada uma revisão abrangente sobre a explicabilidade da inteligência artificial no contexto de aprendizagem profunda. O estudo aborda uma taxonomia detalhada e os principais desafios relacionados à explicabilidade, baseando-se nos princípios do *National Institute of Standards* sobre explicação, significado, precisão e limites do conhecimento. A contribuição significativa desse artigo é a classificação das técnicas de XAI em cinco categorias: explicações*post-hoc*, explicações intrínsecas, explicações transparentes, explicações interativas e explicações híbridas. No contexto da aplicação prática, [Samek and Müller 2019] discutem a importância da interpretabilidade e explicabilidade na IA, especialmente em aplicações críticas como a medicina. Eles argumentam que a falta de interpretabilidade é um problema significativo para as técnicas de aprendizado profundo e também explora a ideia de que a interpretabilidade é um processo contínuo e dinâmico, enfatizando a importância da colaboração entre especialistas em IA e usuários finais para criar modelos precisos e interpretáveis.

Ao longo dos anos, várias técnicas foram propostas para abordar a questão da explicabilidade em modelos de aprendizado de máquina. Por exemplo, o LIME, desenvolvido por [Ribeiro et al. 2016], é uma técnica que visa tornar as previsões de modelos de Aprendizado de Máquina mais compreensíveis. Ele funciona criando uma versão simplificada do modelo complexo para explicar cada previsão individual. Essa versão simplificada, geralmente um modelo linear, permite identificar as características mais importantes que influenciaram a decisão do modelo. O LIME se destaca por sua versatilidade, podendo ser aplicado a uma ampla variedade de modelos. Semelhantemente, [Lundberg and Lee 2017] desenvolveram o SHAP, que utiliza conceitos da teoria dos jogos para atribuir a cada característica um valor que representa sua contribuição para a previsão final. Essa técnica oferece explicações tanto locais (para uma única previsão) quanto globais (para o modelo como um todo), permitindo uma análise mais profunda e consistente.

Nos últimos anos, diversas ferramentas foram desenvolvidas para lidar com a explicabilidade em modelos de aprendizado de máquina, incluindo a IBM *AI Explainability 360* [Arya et al. 2022]. Essa ferramenta oferece uma coleção de algoritmos, interpretadores e visualizações para auxiliar os usuários na compreensão e confiança em modelos de IA. Integrando técnicas de explicabilidade, como *LIME* e *SHAP*, com o objetivo de atender a diferentes necessidades e contextos, proporciona uma plataforma para a análise de modelos de aprendizado de máquina. A InterpretML [Nori et al. 2019] segue uma proposta semelhante, desenvolvida para fornecer explicações transparentes para modelos de aprendizado de máquina. Suportando técnicas como GAM (Generalized Additive Models) e SHAP, o InterpretML permite uma análise detalhada das previsões dos modelos, ajudando os desenvolvedores a compreender melhor os modelos que estão construindo e utilizando. A ferramenta *DALEX (Descriptive mAchine Learning EXplanations)* [Biecek 2018] permite a análise de modelos de aprendizado de máquina, oferecendo abordagens como *Partial Dependence Plots* (PDPs), *Accumulated Local Effects* (ALE) *Plots*, *Break Down Plots*, *Ceteris Paribus Plots* e *Permutational Variable Importance*. Destaca-se por sua capacidade de fornecer explicações detalhadas e acessíveis, facilitando a compreensão das decisões tomadas pelos modelos de aprendizado de máquina.

Ao contrário de abordagens como *LIME* e *SHAP*, que oferecem explicações locais e globais das previsões dos modelos, focando em explicar previsões específicas ou a

contribuição de variáveis individuais, o IMAT se destaca por fornecer uma visualização intuitiva e abrangente de todo o processo de decisão. Ele mapeia detalhadamente cada etapa do processamento de dados dentro dos modelos, permitindo que os desenvolvedores compreendam melhor o funcionamento interno dos modelos e identifiquem possíveis melhorias e ajustes necessários. De acordo com o princípio "overview first, zoom & filter, details-on-demand" [Shneiderman 2003], considera-se uma visualização intuitiva aquela que permite ao usuário compreender a estrutura geral antes de explorar detalhes específicos, sem exigir esforço cognitivo excessivo. No IMAT, esse princípio se materializa da seguinte forma: (i) as camadas da rede são codificadas por elipses cinzas; (ii) cada neurônio é representado por uma elipse azul contendo o cálculo dos pesos e elipses amarelas contendo a adição do *bias*; (iii) a ativação (cor verde) ou inativação (cor cinza) do neurônio; (iv) retângulo vermelho contendo o resultado final. Essa combinação de codificação visual permite ao usuário obter uma visão global imediata e, em seguida, investigar seletivamente as partes mais relevantes do fluxo de informações.

A importância do IMAT reside em sua capacidade de fornecer uma visão completa e compreensível do processo de tomada de decisão dos modelos de aprendizado de máquina, promovendo confiança e uma adoção mais ampla dessas tecnologias em diversos setores. Ferramentas como *IBM AI Explainability 360* e *DALEX* oferecem funcionalidades de interpretabilidade, mas nenhuma combina a capacidade de detalhar visualmente o fluxo de dados com a facilidade de uso que o IMAT proporciona. O IMAT foi desenvolvido não apenas para preencher lacunas identificadas na literatura de XAI, mas também para oferecer uma solução prática e acessível que pode ser aplicada em diversos setores. Ao tornar os modelos de aprendizado de máquina mais interpretáveis, o IMAT facilita a adoção ética e transparente da IA.

4. IMAT

O IMAT[1] é uma ferramenta desenvolvida para oferecer uma análise compreensível e detalhada de modelos de aprendizado de máquina. Atualmente, a ferramenta oferece suporte exclusivo ao algoritmo MLP, focando em fornecer uma visão clara e detalhada do processo de decisão deste tipo de modelo através de uma arquitetura composta por um *backend* implementado em *Flask* e um *frontend* desenvolvido em HTML, CSS e JavaScript.

O *Flask* é um *framework* de *micro web* em Python, que fornece as ferramentas necessárias para criar aplicativos web de forma modular e extensível [Grinberg 2018]. No IMAT, o *Flask* é responsável pelo processamento dos dados, treinamento do modelo e geração dos fluxogramas. O *frontend* oferece uma interface amigável para interação com a ferramenta, permitindo aos usuários carregar conjuntos de dados em formato CSV, configurar parâmetros do modelo, como número de camadas, funções de ativação, otimizadores e funções de perda, além de visualizar os fluxogramas gerados. Utilizando *TensorFlow* que é uma plataforma de código aberto para aprendizado de máquina e *Keras* que é uma API de alto nível que facilita a construção e treinamento de modelos de aprendizado profundo [et al. 2015], o módulo de treinamento de modelos permite o treinamento de modelos MLP com configurações personalizáveis. Assim, o modelo pode ser configurado com diferentes números de camadas e unidades em cada camada, funções de ativação como *ReLU (Rectified Linear Unit)*, *Sigmoid* e *Tanh (Hyperbolic Tangent)*, otimizadores

[1]O código-fonte está disponível em nosso repositório.

como *Adam* e *RMSprop (Root Mean Square Propagation)*, e funções de perda como *MSE (Mean Squared Error)* e *MAE (Mean Absolute Error)*. Durante o treinamento, o modelo é ajustado para minimizar a função de perda escolhida, utilizando o conjunto de dados fornecido pelo usuário.

Para cada camada do MLP, a IMAT extrai as saídas intermediárias, permitindo uma visualização detalhada de como os dados são transformados ao longo da rede. Esse processo envolve a criação de um modelo intermediário que fornece as saídas de cada camada para os dados de entrada, ajudando a compreender o fluxo de informação e as transformações aplicadas a cada etapa. Utilizando a biblioteca de código aberto *Graphviz* para gerar gráficos estruturais e representações em diagramas baseados em descrições textuais [Ellson et al. 2001], a ferramenta gera dois tipos de fluxogramas: detalhado e resumido. O fluxograma detalhado mostra todas as operações matemáticas, pesos, *bias* e ativações em cada camada, enquanto o fluxograma resumido fornece uma visão geral do fluxo de dados pelo modelo, sem incluir todos os cálculos detalhados. Esses fluxogramas são fundamentais para entender o comportamento interno do modelo e as contribuições de cada variável de entrada.

O custo de geração de fluxogramas na IMAT cresce linearmente com o número de pesos representados. Para um MLP com L camadas e N_i neurônios na i-ésima camada, a versão detalhada percorre todos os pesos do modelo, dados por:

$$W = \sum_{i=1}^{L} N_{i-1} N_i,$$

sendo que, para cada peso, são criados nós e arestas correspondentes às operações de multiplicação, soma do viés e aplicação da função de ativação. Esse processo resulta em um custo de tempo e memória proporcional a $O(W)$.

Na versão resumida, a visualização é limitada a K_d neurônios por camada, o que reduz significativamente a complexidade para:

$$O(L \cdot K_d),$$

que se simplifica para $O(L)$ quando K_d é constante, uma vez que o número total de elementos no grafo deixa de depender do tamanho real da rede.

O processo de desenvolvimento da ferramenta seguiu várias etapas, desde o design inicial até a implementação e testes realizados. A Figura 1 apresenta o *design* intuitivo e acessível da IMAT, desenvolvida para ser utilizada por usuários com diferentes níveis de conhecimento em aprendizado de máquina através de uma interface projetada para facilitar a carga de dados, configuração do modelo e visualização dos resultados, garantindo que os usuários possam facilmente configurar e interpretar os modelos gerados pela ferramenta. A ferramenta foi submetida a uma série de testes para confirmar a precisão dos modelos gerados e a clareza dos fluxogramas. Os testes incluíram verificações para confirmar a consistência dos resultados, testes de carga para observar o desempenho do sistema com grandes volumes de dados, e avaliações de usabilidade para certificar que a interface é intuitiva e fácil de navegar. Além disso, foram feitos testes específicos para garantir que a ferramenta pudesse processar grandes conjuntos de dados de maneira eficiente, assegurando a capacidade de resposta do sistema em diferentes condições de uso.

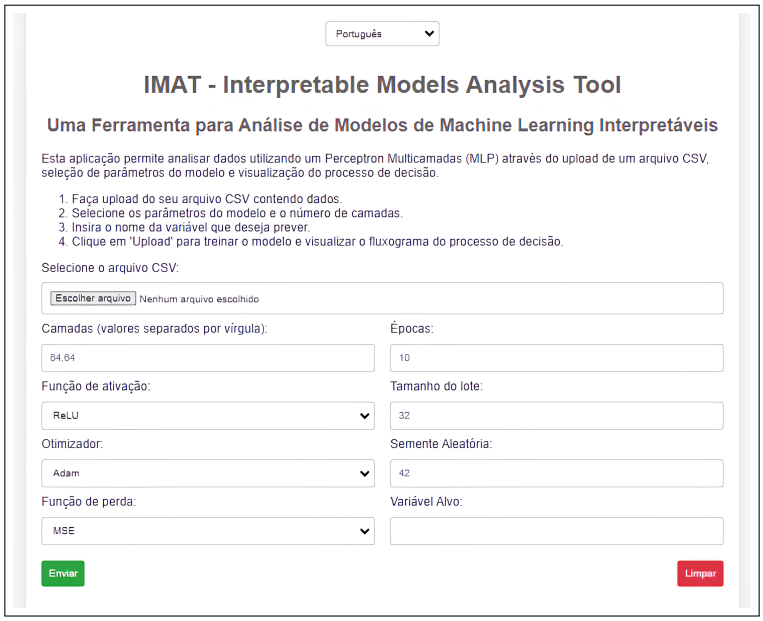

Figure 1. Design da IMAT.

Por fim, o fluxo operacional da IMAT organiza-se em seis etapas sucessivas: (i) o usuário realiza o upload do conjunto de dados em formato CSV; (ii) define, por meio da interface, os hiperparâmetros do MLP (número de camadas e neurônios, funções de ativação, otimizador, taxa de aprendizagem etc.); (iii) a ferramenta executa o treinamento do modelo, acompanhando as métricas e a validação; (iv) ao término do treinamento, são extraídas as ativações e os pesos de cada camada; (v) essas informações são então enviadas ao módulo de visualização, que gera automaticamente um fluxograma detalhado e uma versão resumida com agrupamentos.

5. Estudo de Caso

Nesta seção, apresentamos um estudo de caso que demonstra o uso prático da IMAT através de uma análise detalhada do funcionamento de um modelo de rede neural do tipo MLP aplicado à análise de sentimentos de *tweets*, utilizando o dataset *Sentiment140*. Analisamos exemplos de aplicação da IMAT, comparamos a ferramenta com outras soluções existentes e interpretamos os resultados obtidos, discutindo as implicações das descobertas e as limitações do estudo.

O algoritmo MLP possui uma arquitetura de rede neural *feedforward* composta por multiplas camadas de neurônios. Em termos gerais, o MLP consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio realiza uma operação linear, multiplicando as entradas por pesos e somando um viés, e em seguida aplica uma função de ativação nao linear, como por exemplo a *ReLU*, *Sigmoid* ou *Tanh*, para introduzir complexidade e permitir a modelagem de relações não lineares entre os dados. Durante o treinamento, o algoritmo utiliza o método de retropropagação para ajustar os pesos e vieses, minimizando uma função de perda, o que permite que a rede aprenda representações úteis dos dados de entrada [Taud and Mas 2017].

O conjunto de dados *Sentiment140* utilizado nesse estudo de caso, consiste em *tweets* rotulados, onde os sentimentos são convertidos para uma tarefa de classificação

binária (positivo ou negativo) [Go et al. 2009]. Os textos passaram por um processo de pré-processamento que incluiu limpeza (remoção de URLs, menções, *emojis* e *stopwords*) e a transformação em vetores numéricos utilizando TF-IDF (*Term Frequency-Inverse Document Frequency*) com 1000 atributos. Devido a limitação de processamento, foi utilizada uma amostra reduzida de 1000 registros para a etapa de treinamento. O modelo foi configurado com uma camada oculta de 50 neurônios com função de ativação *ReLU* e uma camada de saída com 1 neurônio e função *Sigmoid*, que gera uma probabilidade representando a confianca na classificação do sentimento. A Figura 2 representa o fluxograma resumido do processo decisório do modelo, onde é possível visualizar as principais etapas de processamento dos dados até a obtenção da previsão final. O modelo completo contêm 50 neurônios na primeira camada, devido a limitação de páginas foi utilizado apenas o fluxograma resumido com 5 neurônios.[2]

O fluxograma ilustra o processo decisório do modelo para identificar o sentimento contido na frase *"Happy birthdayy! You are my hero, keep up the good work!"*. Além da frase original, o nó de entrada exibe a frase vetorizada através das features TF-IDF. Na camada oculta (Hidden Layer 1), que contém 50 neurônios com ativação *ReLU*, o fluxograma detalha os cálculos de alguns neurônios. Por exemplo, para o neurônio 0, o valor linear z_0 é calculado como 1.8391, a partir da combinação dos pesos (como -0.0528, -0.0123, 0.0310, -0.0326 e 0.0508) com a entrada e a soma do viés (-0.0055). Após a aplicação da função *ReLU*, o neurônio gera uma ativação de 1.8391, indicando forte contribuição para o processamento. De forma semelhante, outros neurônios, como o neurônio 1 (com $z_1 = 0.0287$) e o neurônio 2 (com $z_2 = 0.6505$), mostram como pequenas variações nos cálculos podem resultar em ativações baixas ou moderadas. Assim, o fluxograma contribui para a visualização da heterogeneidade na resposta da camada oculta ao receber o mesmo input, permitindo identificar quais neurônios são efetivamente ativados e quais permanecem inativos.

Na camada de saída, composta por um único neurônio com ativação *Sigmoid*, as informações extraídas são consolidadas para gerar a predição final do sentimento. No exemplo apresentado, o fluxograma indica que o modelo classificou o *tweet* como *Positive* com uma confiança de 0.5322. Esse resultado é obtido após o processamento dos dados pela camada oculta, onde os cálculos de multiplicação, soma de viés e ativação são propagados para a camada final. A visualização fornecida pela IMAT contribui significativamente para a compreensão do modelo, pois permite visualizar o fluxo de dados e os cálculos intermediários de forma detalhada. Ao mostrar, por exemplo, como a mesma entrada gera respostas variadas entre os neurônios da camada oculta (com alguns neurônios apresentando ativações robustas e outros resultando em zero), o fluxograma torna evidente a heterogeneidade interna e possibilita a identificação de pontos de melhoria, além de facilitar o diagnóstico de possíveis falhas, assim como a avaliação dos pesos e vieses e a realização de ajustes para aprimorar a performance do modelo, facilitando a interpretação das nuances presentes nas decisões finais.

Para comparar a IMAT com outras ferramentas de explicabilidade, analisamos suas funcionalidades em relação ao LIME e ao SHAP. Embora LIME e SHAP forneçam explicações úteis sobre as previsões de modelos de aprendizado de máquina, a IMAT se destaca por sua capacidade de gerar fluxogramas detalhados que representam o processo

[2]Todas as versões estão disponíveis em nosso repositório.

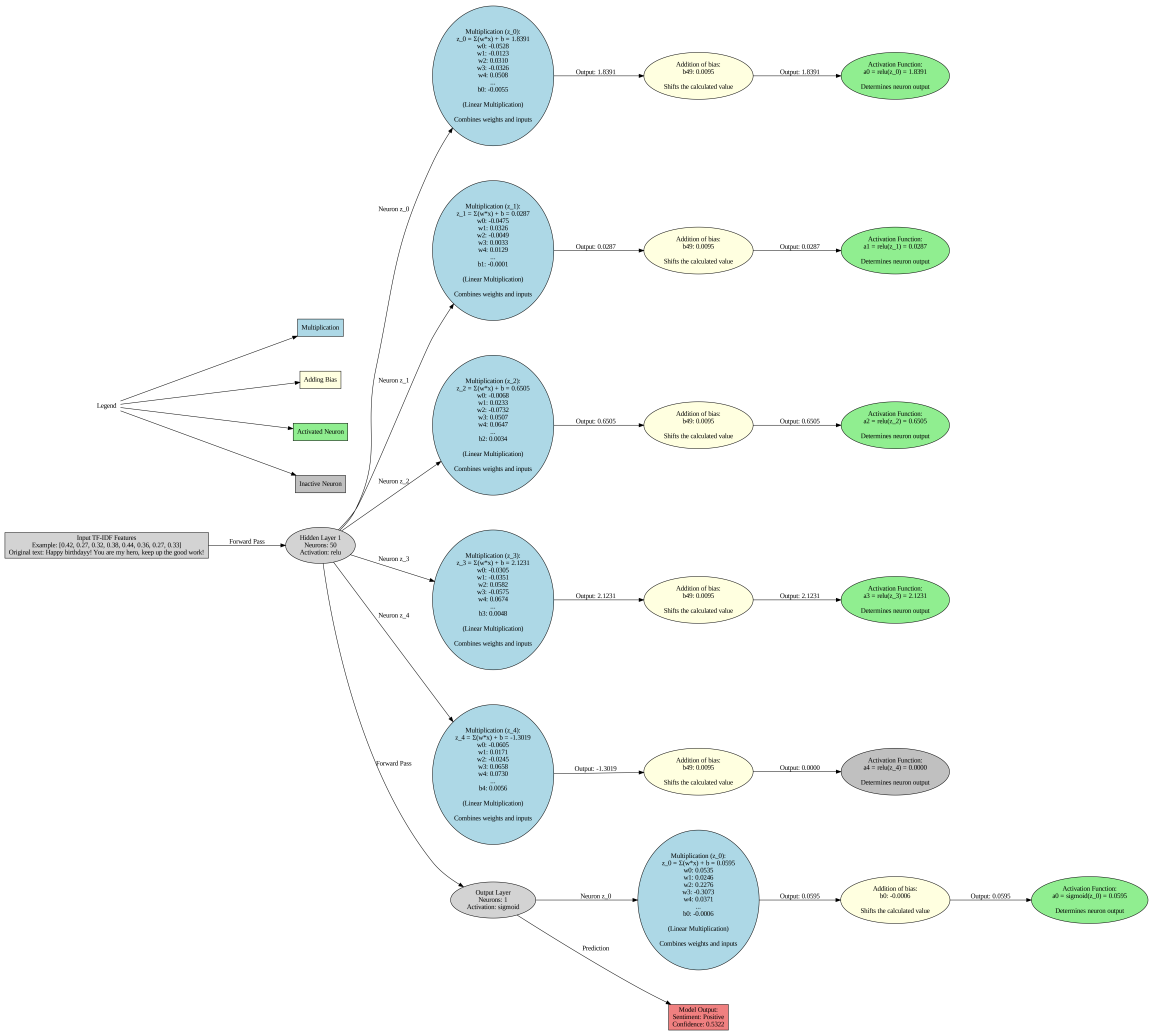

Figure 2. Fluxograma resumido para o conjunto de dados *Sentiment140*.

de decisão completo do modelo. Enquanto LIME e SHAP se concentram em explicações locais e globais das previsões, a IMAT fornece uma visualização intuitiva e abrangente das operações internas do modelo, facilitando a compreensão do fluxo de dados e das transformações aplicadas em cada camada do MLP.

Ao interpretar os resultados obtidos com a IMAT, observamos que a ferramenta facilitou a identificação de características relevantes e o entendimento das contribuições dessas características para as previsões do modelo. Em plataformas como o X, onde as opiniões públicas e as tendências se formam rapidamente, a transparência e a interpretabilidade dos modelos são essenciais para garantir a confiança nas decisões baseadas em dados, como estratégias de marketing, políticas públicas ou campanhas eleitorais. No entanto, o estudo também revelou algumas limitações, pois a IMAT, na versão atual, suporta apenas o algoritmo MLP, limitando sua aplicabilidade a outros tipos de modelos de aprendizado de máquina. Além disso, a complexidade dos fluxogramas gerados pode aumentar significativamente com o crescimento do número de camadas e unidades no modelo, o que pode dificultar a interpretação em modelos muito grandes.

A IMAT demonstrou ser uma ferramenta valiosa para a interpretação de modelos

de MLP, oferecendo visualizações detalhadas que facilitam a compreensão do processo de decisão do modelo, pois quando comparada a outras ferramentas de explicabilidade, a IMAT proporciona uma visão intuitiva e completa das operações internas do modelo, permitindo que os usuários compreendam não apenas as previsões finais, mas também o caminho completo dos dados desde a entrada até a saída. Essa característica é importante em cenários onde a transparência é fundamental, como na área médica, financeira e jurídica. Embora ainda existam oportunidades para expandir suas capacidades, como o suporte a outros tipos de modelos além do MLP, e melhorar a usabilidade em modelos extremamente complexos, a IMAT já se destaca por seu enfoque visual e acessível. Assim, os resultados deste estudo de caso destacam a importância da interpretabilidade na adoção de tecnologias de IA, sugerindo direções futuras para o desenvolvimento de ferramentas de XAI que possam atender a uma gama mais ampla de modelos e aplicações, promovendo uma adoção mais ética e responsável da inteligência artificial em diversas áreas.

6. Conclusões

A IMAT é uma ferramenta desenvolvida para facilitar a interpretação e análise de modelos de aprendizado de máquina, especificamente o MLP. Através de um estudo de caso detalhado, demonstramos a aplicabilidade da IMAT no contexto das redes sociais, através da análise de sentimentos de *tweets*, ilustrando como a ferramenta gera fluxogramas resumidos que representam claramente o processo de tomada de decisão do modelo, mostrando como o modelo identifica e classifica o sentimento presente na mensagem analisada.

O objetivo da IMAT é transformar modelos complexos em representações visuais compreensíveis, facilitando a compreensão e a análise detalhada dos processos internos dos modelos. As figuras geradas pela ferramenta, como demonstrado no exemplo, destacam as principais etapas de processamento, desde a entrada dos dados até a previsão final, promovendo uma maior transparência e confiança nos modelos de aprendizado de máquina. A comparação com outras ferramentas de explicabilidade, como LIME e SHAP, ressaltou a vantagem da IMAT em proporcionar uma visão abrangente e intuitiva das operações internas dos modelos. Assim, a importância da IMAT para a análise de modelos interpretáveis é reafirmada pela sua capacidade de fornecer visualizações detalhadas que não apenas facilitam a compreensão do processo de decisão dos modelos, mas também permitem identificar características relevantes e entender suas contribuições para as previsões, sendo importante em aplicações onde a transparência e a interpretabilidade são essenciais, como nas áreas médica, financeira e jurídica, onde a confiança nos modelos é fundamental para a tomada de decisões informadas e responsáveis.

Embora a IMAT já se apresente como uma ferramenta robusta e eficaz, existem oportunidades para expandir suas capacidades. Trabalhos futuros podem explorar testes com grandes volumes de dados, suporte a outros tipos de modelos de aprendizado de máquina além do MLP, bem como a integração de funcionalidades adicionais que possam melhorar a usabilidade da ferramenta em modelos extremamente complexos. Além disso, a expansão da IMAT para suportar outros modelos de aprendizado profundo, como redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs), visa ampliar seu alcance e aplicabilidade. Adicionalmente, planeja-se desenvolver assistentes automáticos de interpretação para gerenciar a complexidade dos modelos em larga escala.

References

[Adadi and Berrada 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

[Agarwal and Das 2020] Agarwal, N. and Das, S. (2020). Interpretable machine learning tools: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1528–1534. IEEE.

[Alves and ANDRADE 2022] Alves, M. A. S. and ANDRADE, O. d. (2022). Da "caixa-preta" à "caixa de vidro": o uso da explainable artificial intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. *Direito Público*, 18(100).

[Angelov et al. 2021] Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

[Arrieta et al. 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

[Arya et al. 2022] Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., et al. (2022). Ai explainability 360: Impact and design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12651–12657.

[Assis et al. 2023] Assis, A., Véras, D., and Andrade, E. (2023). Explainable artificial intelligence-an analysis of the trade-offs between performance and explainability. In *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE.

[Bhattacharya 2022] Bhattacharya, A. (2022). *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd.

[Biecek 2018] Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, 19(84):1–5.

[Cortiz 2021] Cortiz, D. (2021). Inteligência artificial: conceitos fundamentais. *VAINZOF, Rony; GUTIERREZ, Adriei. Inteligência artificial: sociedade, economia e Estado. São Paulo: Thomson Reuters*, pages 45–60.

[Ellson et al. 2001] Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., and Woodhull, G. (2001). *Graphviz - Graph Visualization Software*. Software available from http://graphviz.org/.

[et al. 2015] et al., M. A. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

[Go et al. 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision.

[Gregor and Benbasat 1999] Gregor, S. and Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530.

[Grinberg 2018] Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. Sebastopol, CA. ISBN 978-1-491-95791-3.

[Kim et al. 2016] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

[Lundberg and Lee 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[Meske and Bunde 2020] Meske, C. and Bunde, E. (2020). Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 54–69. Springer.

[Miller 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

[Molnar 2020] Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

[Nori et al. 2019] Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

[Paes et al. 2022] Paes, V., Araújo, D., Brito, K., and Andrade, E. (2022). Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72, Porto Alegre, RS, Brasil. SBC.

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Salih et al. 2023] Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., and Lekadir, K. (2023). Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv preprint arXiv:2305.02012*.

[Samek and Müller 2019] Samek, W. and Müller, K.-R. (2019). Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22.

[Shneiderman 2003] Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier.

[Silva et al. 2021] Silva, H., Andrade, E., Araújo, D., and Dantas, J. (2021). Sentiment analysis of tweets related to sus before and during covid-19 pandemic. *IEEE Latin America Transactions*, 20(1):6–13.

[Søgaard 2023] Søgaard, A. (2023). On the opacity of deep neural networks. *Canadian Journal of Philosophy*, 53(3):224–239.

[Sundar 2020] Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii). *Journal of Computer-Mediated Communication*, 25(1):74–88.

[Taud and Mas 2017] Taud, H. and Mas, J.-F. (2017). Multilayer perceptron (mlp). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer.