josif 2024
16ª JORNADA CIENTÍFICA E TECNOLÓGICA
13º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
ISSN: 2319-0124
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO CNPq FAPEMIG

UMA ANÁLISE COMPARATIVA ENTRE OS CLASSIFICADORES ADABOOST E NAIVE BAYES UTILIZANDO LOGS DE MOUSE NA IDENTIFICAÇÃO DE WEB BOTS

Fulvio STEFANINE[1]; Taffarel BRANT-RIBEIRO[2]

RESUMO

O avanço da tecnologia proporcionou uma série de benefícios significativos para a sociedade, mas também trouxe novos riscos, especialmente no campo da segurança cibernética. Os bots têm se tornado um problema crescente, realizando atividades prejudiciais na internet e imitando comportamentos humanos. Com cerca de 66% da população mundial conectada à internet, a necessidade de detectar e mitigar essas ameaças se tornou ainda mais crucial. Este estudo examina a eficácia dos algoritmos Naive Bayes e AdaBoost na detecção de bots que simulam movimentos humanos com o mouse. A pesquisa é dividida em várias etapas: a criação de logs de navegação para coleta de dados, o processamento desses dados, a implementação dos algoritmos Naive Bayes e AdaBoost, e a avaliação da eficácia desses métodos na identificação de bots. Como resultados esperados almeja-se que este projeto a detecção de comportamentos automatizados, proporcionando uma abordagem mais eficaz para enfrentar as ameaças cibernéticas e proteger sistemas online contra atividades fraudulentas.

Palavras-chave: Avanço da tecnologia; Segurança cibernética; Comportamentos automatizados; Mitigação de ameaças; Efetividade de algoritmos .

1. INTRODUÇÃO

A crescente dependência da internet pela sociedade moderna trouxe significativos benefícios, como acesso rápido à informação, automação de tarefas, e oportunidades de emprego (Rahman; Tomar, 2020). No entanto, essa dependência também amplificou problemas relacionados à segurança digital. De acordo com Richabadas (2023), aproximadamente 50% do tráfego digital é gerado por bots, dos quais 30% são maliciosos.

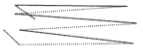
Bots são programas automatizados que realizam tarefas repetitivas na web, como indexação e extração de dados (Iliou et al., 2021). Embora muitos bots sejam benéficos, há uma crescente preocupação com bots maliciosos que simulam comportamento humano para evitar detecção (Rahman; Tomar, 2020).

Esses bots podem realizar atividades prejudiciais, desde cliques automatizados até o roubo de informações pessoais (Iliou et al., 2021). Dado o impacto desses bots, técnicas de Aprendizado

[1]Estudante, IFSULDEMINAS - *Campus* Passos. E-mail: fulvio.junior@alunos.ifsuldeminas.edu.br
[2]Orientador, IFSULDEMINAS – *Campus* Passos. E-mail: brant.ribeiro@ifsuldeminas.edu.br

de Máquina (ML) têm sido empregadas para detectar e mitigar essas ameaças sem afetar usuários legítimos (Walt; Eloff, 2018).

Tabela 1: Exemplo de comportamentos de humanos, bots moderados e bots avançados.

	Humano	**Bot Moderado**	**Bot Avançado**
Características	Movimentos realizados por usuários. Movimentos de indecisão e inconstância.	Seleção aleatória de hiperlink. Movimentos diretos e precisos.	Seleção heurística de hiperlink. Movimentos avançados e precisos.
Exemplo 1			
Exemplo 2			

Fonte: Adaptado de Iliou et al. (2021).

De acordo com Iliou (2021), bots moderados possuem uma impressão digital semelhante à de um navegador, mas não conseguem reproduzir comportamentos humanos. Por outro lado, bots avançados não apenas replicam a impressão digital de um navegador, mas também simulam comportamentos humanos, tornando-se mais difíceis de serem detectados.

2. FUNDAMENTAÇÃO TEÓRICA

A Inteligência Artificial (IA) é um dos ramos da computação onde são realizados inúmeros estudos relacionados ao desenvolvimento sem intervenção humana. Já o aprendizado de máquina, fazendo parte da Inteligência Artificial, possui capacidade para interpretar dados sensoriais, compreender e responder à linguagem humana de forma natural e eficiente (Russell; Norvig, 2020).

Dito isso, utilizar desta ferramenta para análise de logs de navegação e movimentos de mouse é crucial para entender o comportamento dos usuários e melhorar a segurança online. Os logs de navegação contêm informações como links visitados e horários de acesso, enquanto os movimentos de mouse oferecem dados sobre a interação do usuário com a interface (Zhou et al., 2020).

Segundo Wang (2021), algoritmos como o AdaBoost melhoram a precisão combinando classificadores fracos e ajustando seus pesos com base no desempenho. Como também o Naive Bayes que utiliza o teorema de Bayes para calcular probabilidades de classificação, assumindo a independência entre características (Ma; Yamamori; Thida, 2020) podem ser úteis no processo de compreensão e classificação dos dados..

3. MATERIAIS E MÉTODOS

O dataset utilizado será o criado por Yıldırım (2021), este dataset inclui variáveis essenciais para a implementação e desenvolvimento de bots. Os logs dos usuários humanos foram registrados através de ferramentas de captura de eventos do navegador, que coletaram dados detalhados, como coordenadas do cursor, cliques e períodos de inatividade (Tabela 2).

Tabela 2: Descrição de colunas detalhadas presentes no dataset.

Variável	**Tipo**	**Descrição**
Tipo de ação	Categórico	Movimento ou clique
Data/Hora	Float	Tempo de movimento em segundos
X	Numérico	Posição X do cursor do mouse
Y	Numérico	Posição Y do cursor do mouse
Botão	Categórico	Botão esquerdo/direito ou nenhum caso esteja em movimento
Estado	Categórico	Pressionado, solto ou movimento

Fonte: Yıldırım (2021).

Bots serão desenvolvidos com objetivo de recriar as variáveis dos usuários reais para aderir ao dataset existente. O pré-processamento removerá dados irrelevantes e redundantes para garantir consistência na segmentação dos dados para treinamento, validação e teste.

O conjunto de treinamento (70% dos dados) será usado para treinar o modelo, enquanto o conjunto de validação (15%) ajustará hiperparâmetros e prevenirá overfitting. O conjunto de teste (15%) avaliará a precisão do modelo. Os algoritmos Naive Bayes e AdaBoost serão implementados em Python com Scikit-learn, e a avaliação será feita usando métricas de Acurácia, Precisão, Revocação e Medida-F1.

4. RESULTADOS ESPERADOS E CONTRIBUIÇÕES

Dado o objetivo deste projeto, espera-se que a implementação e análise dos algoritmos Naive Bayes e AdaBoost para a detecção de web bots, com base nos logs de movimento do mouse,

resultem em uma estrutura robusta e confiável. Além disso, espera-se que as análises comparativas detalhadas entre os resultados de cada algoritmo ajudem a identificar a técnica mais eficaz na detecção e generalização para diferentes tipos de bots.

Os resultados obtidos deverão fornecer informações relevantes sobre a eficácia dos algoritmos utilizados, visando incentivar novos métodos de detecção com o objetivo de mitigar atividades maliciosas em ambientes web. Em suma, os resultados poderão servir de base para o desenvolvimento de ferramentas práticas e soluções robustas, além de fornecer uma referência para estudos futuros.

REFERÊNCIAS

ILIOU, C. et al. Detection of Advanced Web Bots by Combining Web Logs with Mouse Behavioural Biometrics. **Association for Computing Machinery**, s/l, v. 2, n. 3, p. 1–26. 2021.

MA, T.; YAMAMORI, K.; THIDA, A. A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification. In: IEEE 9TH GLOBAL CONFERENCE ON CONSUMER ELECTRONICS (GCCE), 2020, Kobe, Japão, IEEE, p. 324-326.

RAHMAN, R; TOMAR, D. A new web forensic framework for bot crime investigation. **Forensic Science International Digital Investigation**, s/l, v. 33, p. 300943, 2020.

RICHABADAS, T. **Threat Spotlight: How Bad Bot Traffic Is Changing**. Disponível em: <https://blog.barracuda.com/2023/10/18/threat-spotlight-bad-bot-traffic-changing#:~:text=From%20January%202023%20to%20June,up%2039%25%20of%20internet%20traffic>. Acesso em abr. 2024.

RUSSELL, S.; NORVIG, P. Inteligência Artificial: A Modern Approach. 4. ed. Inglaterra: Pearson, 2020.

WANG, W; SUN, D. The improved AdaBoost algorithms for imbalanced data classification. **Information Sciences**, s/l, v. 563, p. 358-374, 2021.

WALT, E. V. D; ELOFF, J. Using Machine Learning to Detect Fake Identities: Bots vs Humans**. IEEE Access**, s/l, v. 6, n. 1, p. 6540-6549.

YILDIRIM, Metehan; KILIÇ, Arjen Aykan; ANARIM, Emin. Boğaziçi University Mouse Dynamics Dataset**. Journal Data in Brief**, v. 2, 2021.

ZHOU, Z. et al. Analyzing Web Navigation Logs for User Behavior Insights**. Journal of Internet Services and Applications**, s/l, v. 11, n. 3, p. 45-60, 2020.