INSTITUTO BRASILEIRO DE ENSINO, DESENVOLVIMENTO E PESQUISA - IDP
CAMPUS ASA NORTE

DESENVOLVIMENTO DE UMA FERRAMENTA PARA O MONITORAMENTO PERIÓDICO DE PLATAFORMAS DIGITAIS DE APOSTAS

Colocar os nomes de vocês aqui, em letra maiúscula, um abaixo do outro.
MATHEUS NILTON BIOLOWONS
RAFAEL PEREIRA CÂNDIDO
GUILHERME AUGUSTO MONTALVÃO

BRASÍLIA - DF
2024

SUMÁRIO

1. INTRODUÇÃO

A crescente popularidade das apostas de quota fixa no Brasil e no cenário internacional, impulsionada por um mercado em rápida expansão e por ampla publicidade em canais televisivos, redes sociais e plataformas online, tem demandado maior atenção regulatória e mecanismos de fiscalização. A disseminação dessas apostas, sobretudo entre o público jovem, aliada ao envolvimento de operadores licenciados em jurisdições estrangeiras e à ainda incipiente regulamentação nacional, cria um ambiente complexo no qual o dever de informar o consumidor sobre restrições etárias e riscos associados ao jogo torna-se um ponto central.

Nesse contexto, a necessidade de aferir o cumprimento das normas legais e dos padrões de autorregulamentação publicitária sobre a apresentação de cláusulas de advertência nos sítios eletrônicos dos operadores de apostas de quota fixa motivou o desenvolvimento de um software específico. Esse software, derivado da pesquisa acadêmica de mestrado que investiga a conformidade dos operadores com a legislação brasileira, foi concebido para verificar, de forma sistemática e padronizada, a presença das informações obrigatórias na página de abertura desses sítios.

O presente relatório técnico descreve o processo de concepção, desenvolvimento e avaliação do software. Primeiramente, apresenta-se o objetivo da aplicação: automatizar a verificação dos sítios de operadores de apostas, identificando a presença das cláusulas de advertência sobre restrição etária e riscos associados à atividade. Em seguida, expõem-se os fundamentos teóricos das tecnologias adotadas, incluindo as linguagens de programação, bibliotecas e ferramentas que embasam a análise. A abordagem metodológica é detalhada, abrangendo o levantamento de requisitos, a modelagem do sistema, a definição de critérios de busca e reconhecimento de padrões, bem como a elaboração do fluxograma que ilustra o fluxo de processamento.

O relatório também discute o desenvolvimento do software em si, descrevendo a arquitetura interna, as principais rotinas, testes realizados e o fluxo lógico da aplicação. Serão abordadas, além disso, as limitações encontradas, sejam decorrentes das restrições técnicas ou das especificidades do ambiente regulatório e legal, bem como sugestões para aprimoramentos futuros. Por fim, apresenta-se uma seção de considerações finais, na qual se destacam as contribuições do software para a análise da conformidade dos operadores de apostas às normas vigentes, bem como as perspectivas de evolução da ferramenta,

considerando o avanço da regulamentação nacional e o aperfeiçoamento constante das tecnologias empregadas.

2. OBJETIVO

O objetivo principal do desenvolvimento do software é assegurar a verificação da presença e da adequação das cláusulas de advertência de risco e de restrição etária, conforme exigido pela legislação vigente para os sítios eletrônicos de operadores de apostas. Essa verificação busca garantir que essas cláusulas estejam não apenas devidamente incluídas nas plataformas, mas também apresentadas de forma clara e visível, permitindo que sejam facilmente percebidas pelos apostadores, promovendo, assim, maior conformidade com as normas regulatórias e a proteção dos usuários. Nesse sentido, o programa busca identificar, de forma automática, se as expressões obrigatórias de restrição etária e de riscos associados à atividade de aposta estão fielmente reproduzidas e localizadas em área de fácil percepção, permitindo avaliar o cumprimento do dever legal de informação, conforme as normas aplicáveis à proteção e defesa do consumidor.

Essa aferição reveste-se de especial importância porque atende à exigência legal de assegurar que o consumidor receba informações acerca dos riscos associados à atividade de apostas e da restrição etária estabelecida. Ao verificar não apenas a presença, mas também a localização exata das cláusulas de advertência, é possível constatar se os operadores estão em conformidade com a regulamentação aplicável.

A presença das cláusulas visa garantir que o consumidor tenha acesso imediato e inequívoco aos alertas necessários, antes mesmo de interagir mais profundamente com o conteúdo do sítio eletrônico. Essa prática contribui para evitar ambiguidades, minimizar o potencial de indução ao erro e reforçar o cumprimento dos deveres de informação previstos na legislação consumerista e específica, refletindo um compromisso com a proteção e a conscientização do público frente a uma atividade inerentemente arriscada.

3. FUNDAMENTAÇÃO DAS TECNOLOGIAS UTILIZADAS

Esta seção apresenta os conceitos e as tecnologias utilizados para a construção da aplicação proposta, que realiza a coleta e a análise automatizada de informações sobre os sítios eletrônicos dos sítios eletrônicos de apostas. Serão discutidos os fundamentos do OCR

(Reconhecimento Ótico de Caracteres), técnicas de tratamento de dados (Web Scraping), agendamento de tarefas periódicas (Cron Jobs) e a aplicação de estruturas de dados em filas para o gerenciamento e processamento de tarefas assíncronas.

3.1 OCR (RECONHECIMENTO ÓTICO DE CARACTERES)

O Reconhecimento Ótico de Caracteres (OCR) é uma técnica amplamente utilizada para converter informações visuais em dados processáveis por sistemas computacionais, sendo amplamente empregado em cenários como a digitalização de documentos físicos para facilitar seu arquivamento e busca eletrônica. Conforme destacado por C. Patel, A. Patel e D. Patel (2012), o OCR é uma ferramenta indispensável em sistemas que necessitam da extração textual de imagens complexas.

O funcionamento do OCR baseia-se na fragmentação das imagens em blocos de texto e no reconhecimento de padrões para identificar caracteres. Ferramentas como o Tesseract, por exemplo, utilizam redes neurais treinadas para detectar caracteres em diversas fontes e formatações, viabilizando a extração de informações textuais. No contexto das apostas, essa tecnologia desempenha um papel crucial ao identificar avisos obrigatórios inseridos em páginas da web e banners publicitários. Essa funcionalidade é indispensável para verificar o cumprimento dos requisitos mínimos estabelecidos pelo Anexo X do Conselho Nacional de Autorregulamentação Publicitária (CONAR).

Apesar de sua reconhecida agilidade e eficiência, o Reconhecimento Óptico de Caracteres enfrenta limitações que podem comprometer sua precisão. Entre os principais desafios estão a baixa resolução das imagens, o uso de fontes tipográficas não convencionais e a presença de elementos gráficos que interferem na leitura. No entanto, grande parte dessas limitações pode ser mitigada por meio da aplicação de técnicas de pré-processamento de imagens.

3.2 WEB SCRAPING

O Web Scraping é uma técnica automatizada utilizada para a extração de dados em páginas web, convertendo as informações disponíveis em formato visual (imagens) ou textos que podem ser analisados previamente, conforme citado por Khder (2021). Tal abordagem é amplamente utilizada em ambientes onde os dados são apresentados publicamente, mas não são facilmente acessíveis por APIs ou Bancos de dados.

No contexto deste projeto, a aplicação utiliza a biblioteca Puppeteer, baseada em Node.js, que permite a navegação automatizada em páginas web. A utilização dessa ferramenta no monitoramento dos sítios eletrônicos de apostas se faz essencial para a coleta periódica dos dados.

Contudo, a utilização de Web Scraping é bastante trabalhosa, uma vez que as páginas têm uma constante atualização nos seus dados, interfaces, dimensões e até medidas anti-scraping, como CAPTCHAs e/ou restrições de IPs.

3.3 FILAS (QUEUES)

A utilização de filas possui um papel essencial na organização e no processamento das tarefas de scrapping automatizadas. A fila trabalha com os elementos no modelo de FIFO (First In, First Out), onde as primeiras tarefas ao serem alocadas, serão as primeiras a serem processadas. Tal modelo é extremamente útil em sistemas onde ocorrem monitoramento contínuo, assegurando que os itens mais antigos da fila sejam processados antes dos mais recentes, evitando acúmulos desnecessários ou priorizações incorretas de elementos.

Segundo Indra e Sarjono (2010), o modelo FIFO é amplamente utilizado em sistemas de filas para organizar o fluxo de trabalho de maneira eficiente, assegurando que os elementos sejam processados na ordem de chegada. Esse conceito é essencial no contexto do projeto, em que as filas organizam alvos coletados pelo Web Scraper para execução ordenada e sem conflitos das análises dos sítios eletrônicos de apostas.

3.4 CRON JOB

Cron Jobs são tarefas agendadas que permitem a execução de processos, tarefas e comandos automaticamente em intervalos de tempos bem definidos, sendo uma ferramenta extremamente essencial em sistemas distribuídos em que precisam de repetição em tarefas, como definido por Keller (1999). No contexto deste projeto, os Cron Jobs são utilizados para agendar e controlar a execução do Web Scraper, que fará a raspagem dos dados dos sítios eletrônicos de forma contínua e sistemática.

A principal vantagem de utilizar essa abordagem é a possibilidade de manter o sistema automatizado, com consultas periódicas às plataformas de apostas, otimizando a eficiência do procedimento executado pelo Web Scraper e oferecendo uma maior resistência contra sistemas anti-scrapping.

4. METODOLOGIA

A metodologia utilizada consistiu, primeiramente, na identificação de um conjunto de operadores de apostas autorizados, em nível nacional e estadual, e na consolidação dos respectivos domínios eletrônicos em um rol abrangente. Foi utilizada uma lista inicial de 100 empresas que solicitaram autorização ao Ministério da Fazenda até 17 de setembro de 2024. Essas empresas, detentoras de marcas e domínios registrados, estariam autorizadas a oferecer apostas de quota fixa em âmbito nacional durante o período de adequação, até 31 de dezembro de 2024, conforme a legislação vigente (Lei nº 14.790/2023, Portaria SPA/MF nº 827/2024 e Portaria SPA/MF nº1.475/2024)[1].

Além disso, considerou-se um conjunto adicional de 26 empresas com autorização concedida por Estados para a exploração de apostas de quota fixa em seus territórios, de acordo com o art. 35-A da Lei nº13.756/2018. Suas marcas e domínios foram comunicados ao Ministério da Fazenda até 18 de outubro de 2024, em conformidade com a Portaria SPA/MF nº 1.475/2024[2].

Nas referidas listas, constatou-se que algumas empresas apresentaram mais de uma marca, cada qual com domínios próprios, outras obtiveram autorizações tanto em âmbito nacional quanto estadual, e houve ainda aquelas que, embora autorizadas, não informaram domínios de internet por estarem em fase de implantação. Após consolidar as relações das pessoas jurídicas, nacionais e estaduais, e eliminar domínios duplicados, obteve-se um total de 236 endereços de sítios eletrônicos.

As expressões-chave empregadas para a verificação foram extraídas das disposições legais e regulatórias aplicáveis. Em relação à restrição etária, utilizou-se as expressões “18+” e “proibido para menores de 18 anos”, conforme previsto no artigo 13, inciso I, da Portaria SPA/MF nº1.231/2024. Sobre a utilização dessas expressões, observou-se a necessidade de fidelidade literal ao texto previsto nas normas.

---

[1] Lista de empresas que podem ofertar apostas de quota fixa em nível nacional, publicada em 01 out. 2024, atualizada em 18 out. 2024. Disponível em: https://www.gov.br/fazenda/pt-br/composicao/orgaos/secretaria-de-premios-e-apostas/lista-de-empresas/nacionais-18-10.pdf. Acesso em: 28 out. 2024.

[2] Lista de empresas que podem ofertar apostas de quota fixa em nível estadual, publicada em 01 out. 2024, atualizada em 18 out. 2024. Disponível em: https://www.gov.br/fazenda/pt-br/composicao/orgaos/secretaria-de-premios-e-apostas/lista-de-empresas/estaduais-18-10.pdf. Acesso em: 28 out. 2024.

Nesse sentido, não se admitiu qualquer variação semântica ou emprego de sinônimos, visando manter a padronização e a clareza exigidas pelas disposições legais e regulamentares. Assim, expressões como “+18” ou “site para maiores de idade” não foram consideradas equivalentes às formas prescritas (“18+” e “proibido para menores de 18 anos”).

Já no que se refere às advertências sobre avisos de riscos associados às apostas de quota fixa, foram adotadas as frases estabelecidas no item 6 do Anexo X do Código Brasileiro de Autorregulamentação Publicitária, quais sejam: “jogue com responsabilidade”, “apostas são atividades com riscos de perdas financeiras”, “apostar pode levar à perda de dinheiro”, “as chances são de que você está prestes a perder”, “aposta não é investimento”, “apostar pode causar dependência”, “apostas esportivas: pratique o jogo seguro”, “apostar não deixa ninguém rico”, “saiba quando apostar e quando parar” e “aposta é assunto para adultos”.

Em seguida, foi desenvolvido e empregado um software para analisar a página de abertura (página inicial) de cada sítio eletrônico, procedendo à verificação da existência das cláusulas de advertência legalmente exigidas, tanto no código-fonte (linguagem de marcação) quanto em eventuais elementos visuais (imagens). Além disso, a ferramenta examinou a localização dessas informações, avaliando se estavam dispostas na área imediatamente visível (renderizada) ou em sessões acessíveis apenas mediante rolagem da página.

Caso o software identifique mais de uma cláusula de advertência do mesmo tipo na página de abertura (página inicial) do sítio eletrônico do operador de apostas, ele foi programado para priorizar o registro da posição daquela que estiver disposta na área renderizada da página (ou seja, na porção visível sem necessidade de rolagem).

Assim, a metodologia combina a seleção criteriosa da amostra (pessoas jurídicas autorizadas), a preparação tecnológica (desenvolvimento do software) e a análise automatizada do conteúdo e da disposição das advertências, permitindo uma aferição objetiva e padronizada do cumprimento das obrigações informacionais.

5. DESENVOLVIMENTO DO SOFTWARE

5.1 VISÃO GERAL

O desenvolvimento da aplicação foi estruturado em etapas, visando criar uma ferramenta funcional e eficiente ao monitoramento contínuo dos sítios eletrônicos de apostas. A metodologia adotada buscou equilibrar simplicidade e eficiência, utilizando ferramentas modernas que garantem a escalabilidade e a fácil manutenção do sistema.

Inicialmente, seguimos uma abordagem iterativa, o que nos permitiu avançar progressivamente no desenvolvimento e ajustar a aplicação à medida que os testes identificaram pontos de melhoria. Por ser um projeto focado em resolver um problema real e bem delimitado, priorizamos a entrega de resultados práticos em cada etapa.

Para sustentar o fluxo de processamento da aplicação, foram selecionadas tecnologias amplamente utilizadas no contexto atual da indústria de desenvolvimento de software. Entre as ferramentas adotadas destacam-se o Fastify, para a construção da API conhecido por sua leveza e alto desempenho, o MySQL em conjunto com o TypeORM para realizar a ponte entre a aplicação e a persistência dos dados. Tal combinação permitiu que os dados fossem tratados eficientemente e que o banco de dados fosse escalável para atender às necessidades do sistema. O Redis aliado à biblioteca Bull, para execução assíncrono dos processos e gerenciamento de filas. A escolha dessas tecnologias foi fundamentada na facilidade de implementação, robustez e no amplo suporte oferecido pela comunidade de desenvolvedores.

O principal objetivo da aplicação é garantir a execução autônoma dos processos, assegurando a coleta e análise de dados de forma eficiente e precisa. Adicionalmente, a arquitetura do sistema foi projetada considerando aspectos de escalabilidade e elasticidade, possibilitando futuras expansões tanto na aplicação quanto na infraestrutura subjacente, atendendo assim a demandas crescentes de forma sustentável e eficiente.

5.2 DESENVOLVIMENTO E TECNOLOGIAS

O desenvolvimento da aplicação seguiu uma estrutura, focada em garantir um sistema eficiente, escalável e de fácil manutenção. Foi utilizado TypeScript como linguagem principal para assegurar um código mais seguro e legível, aproveitando os benefícios da tipagem estática para minimizar erros. Além disso, utilizou-se o Bun como runtime para a execução do projeto, devido ao seu desempenho superior e inicialização rápida comprado ao runtime do Node.js.

No contexto do processamento assíncrono, o Redis foi utilizado como banco de dados para o gerenciamento de filas, em conjunto com a biblioteca Bull, proporcionando uma organização eficiente e ordenada das tarefas executadas pelo Web Scraper, mesmo em cenários de alta concorrência. Essa abordagem permitiu que o sistema mantivesse um desempenho consistente, mesmo diante de grandes volumes de dados e solicitações simultâneas.

A implementação do Web Scraper foi realizada utilizando a biblioteca Puppeteer, configurada para acessar periodicamente os sites de apostas por meio de agendamentos automatizados (Cron Jobs). O sistema é responsável por identificar e capturar todas as imagens e expressões-chave presentes no código HTML das páginas. Esses elementos, sejam de texto ou imagem, são analisados e submetidos a tratamentos que permitem a identificação e coleta de informações relevantes, como contraste em conteúdos presentes no HTML que não sejam imagens, proporção, distância em relação ao topo da pagina e verificação de sua localização (Viewport).

As imagens coletadas passam, inicialmente, por um processo de sanitização, que compreende a aplicação de técnicas destinadas a melhorar sua qualidade e adequação para análise. Esse processo inclui a aplicar tons de cinza redução de ruídos, o ajuste de contraste e o redimensionamento de imagens no formato PNG com dimensões superiores a 1024 píxeis e de imagens no formato SVG com dimensões maiores que 512 píxeis, tanto na vertical quanto na horizontal, para garantir uma maior precisão na análise das informações. Após essa etapa, as imagens são submetidas a um sistema de Reconhecimento Óptico de Caracteres (OCR), responsável pela extração dos textos contidos nos elementos visuais previamente processados.

Para o desenvolvimento do projeto, foi empregado o modelo avançado de reconhecimento óptico de caracteres (OCR), denominado PaddleOCR, que incorpora uma arquitetura de última geração otimizada para o processamento de documentos complexos e variados. Essa solução é projetada para lidar com textos em diferentes contextos, desde conteúdos simples até formatos sofisticados, como tabelas, fórmulas matemáticas e textos rotacionados em até 180 graus.

O PaddleOCR destaca-se por sua capacidade de lidar com alta densidade de caracteres, múltiplos idiomas e cenários de reconhecimento desafiadores. Além disso, sua arquitetura moderna utiliza módulos avançados, como redes baseadas em transformadores, para otimizar a extração de informações contextuais em imagens de linhas de texto, eliminando a necessidade de RNNs. A integração de estratégias inovadoras, como o treinamento orientado por CTC. O uso de dados aumentados (TextConAug) e modelos pré-treinados com alta-supervisão (TextRotNet), amplia sua eficácia. Adicionalmente, recursos como o UDML (Unified Deep Mutual Learning) e o UIM (Unlabeled Images Mining) aceleram o desempenho e aprimoram os resultados.

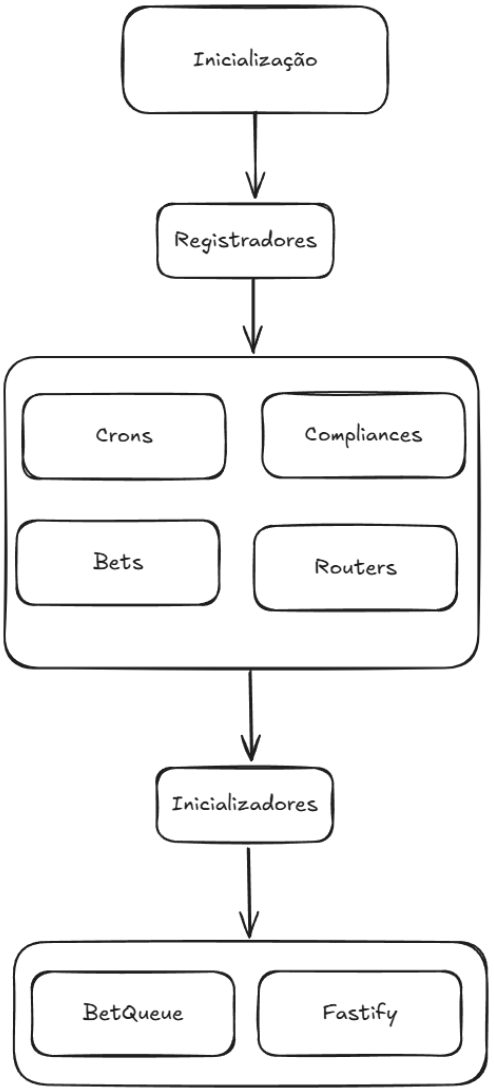
Os textos extraídos passam por um processo de normalização, no qual são aplicadas transformações para padronização, como a remoção de espaços e acentos, além da conversão de todos os caracteres para letras minúsculas. Esse procedimento visa facilitar a comparação com expressões-chave previamente definidas. Após a normalização, para encontrar as cláusulas de aviso de risco e restrição etária os textos das imagens são submetidos a um sistema de análise de similaridade, que permite reconhecer expressões mesmo que incompletas ou parcialmente legíveis. Por exemplo, uma expressão como “ogue com resp0nsabilidade” pode ser corretamente interpretada como “Jogue com Responsabilidade”. Essa integração entre o processo de normalização textual e o sistema de similaridade possibilita verificar se os sítios eletrônicos de apostas analisados estão em conformidade com as normas estabelecidas pela legislação vigente.

Além disso, a aplicação conta com funcionalidades que permitem exportar os dados analisados em formato XLSX (Excel), utilizando a biblioteca ExcelJS. Essa funcionalidade foi pensada para tornar mais simples e acessível à criação de relatórios detalhados, possibilitando que as informações obtidas sejam organizadas de maneira clara e visualmente estruturada. Dessa forma, os resultados podem ser facilmente compartilhados e utilizados por diferentes públicos, garantindo maior transparência e utilidade prática dos dados processados.

O software foi desenvolvido para atuar como uma API para um front-end final, permitindo que todos os processos automatizados sejam executados por meio de requisições HTTP. O termo API (Application Programming Interface) refere-se a uma interface que permite a comunicação entre diferentes sistemas ou componentes de software. Essa comunicação é fundamental no desenvolvimento de soluções modulares e flexíveis, possibilitando que funcionalidades específicas sejam implementadas de forma independente e reutilizável, promovendo maior eficiência e escalabilidade no desenvolvimento.

A integração por meio de uma API permite que o front-end funcione de maneira independente do back-end, simplificando a manutenção e a atualização do software. Além disso, promove a criação de ferramentas modulares, onde diferentes partes do sistema podem ser desenvolvidas, testadas e aprimoradas separadamente. Esse modelo de desenvolvimento modular facilita não apenas a adição de novas funcionalidades, mas também a adaptação da ferramenta a diferentes contextos e plataformas, tornando-a mais versátil e de fácil uso.

5.3 FLUXO DA APLICAÇÃO



5.3.1 REGISTRADORES

Os registradores implementados no sistema desempenham papéis fundamentais no gerenciamento de tarefas, dados e rotas, assegurando a funcionalidade e a organização das operações. Esses componentes foram projetados para executar, de maneira eficiente, a escrita e validação de dados previamente definidos no banco de dados MySQL usando o ORM TypeORM, garantindo consistência e integridade das informações.

5.3.1.1 BETS

Previamente, foram registradas 236 sítios eletrônicos de apostas no banco de dados, utilizando um conjunto de dados (dataset) datado de 18 de outubro de 2024, fornecido em formato CSV. Esse registro foi estruturado de maneira a garantir que as informações estejam prontamente acessíveis para o processamento e análise pelo sistema. Além disso, a aplicação oferece funcionalidades via API que permitem a adição e remoção dos sítios eletrônicos de apostas eficientemente, assegurando flexibilidade na gestão dos dados e possibilitando a adequação às necessidades específicas do usuário ou às atualizações requeridas pelo sistema.

5.3.1.2 CRONS

Os valores predefinidos para a execução de tarefas agendadas são armazenados no banco de dados, facilitando sua reutilização em processos automatizados e garantindo maior eficiência no gerenciamento das atividades programadas. Um exemplo de configuração é o cron `0 */6 * * *`, que define uma tarefa para ser executada a cada seis horas.

No neste contexto, cada casa de aposta é associada a uma configuração de cron específica, permitindo o agendamento automatizado de suas tarefas. Para OS sítios de operadores de apostas previamente registrados no sistema, foi estabelecido um cron com intervalo de execução a cada 12 horas. Esse procedimento assegura que os dados das plataformas sejam coletados e analisados periodicamente, promovendo um monitoramento contínuo e eficiente.

5.3.1.3 CONFORMIDADES (COMPLIANCES)

Palavras e expressões de interesse relacionadas a cláusulas de restrição etária, como “18+” e “proibido para menores de 18 anos”, bem como termos associados aos avisos de riscos, tais como “jogue com responsabilidade”, “apostas são atividades com riscos de perdas financeiras”, “apostar pode levar à perda de dinheiro”, “as chances são de que você está prestes a perder”, “aposta não é investimento”, “apostar pode causar dependência”, “apostas esportivas: pratique o jogo seguro”, “apostar não deixa ninguém rico”, “saiba quando apostar e quando parar” e “aposta é assunto para adultos”, são previamente registradas no banco de dados.

Esses registros desempenham um papel fundamental na análise das plataformas, sendo amplamente utilizados em correlações com tabelas de propriedades de HTML e OCR. Esse procedimento permite o tratamento eficiente dessas informações e a extração precisa de dados, auxiliando diretamente no monitoramento do comprimento das regulamentações aplicáveis. A padronização e o registro de tais expressões garantem maior precisão nos processos automatizados de identificação e validação, contribuindo para a conformidade das plataformas com as normas vigentes.

5.3.1.4 ROTEADORES (ROUTERS)

Para otimizar o uso do framework Fastify, foi implementado um sistema de rotas dinâmicas, inspirado no modelo do Next.js. Nesse sistema, nomes específicos de arquivos determinam o comportamento das rotas. Por exemplo, um arquivo denominado /bets/index.ts é automaticamente registrado como a rota /bets e gerenciado em memória antes de ser repassado para o Fastify.

5.3.2 INICIALIZADORES

Os inicializadores são responsáveis pelo agendamento de tarefas automatizadas com base na configuração precisa de crons. Essa funcionalidade permite a execução programada de processos críticos, promovendo eficiência e reduzindo a necessidade de intervenções manuais.

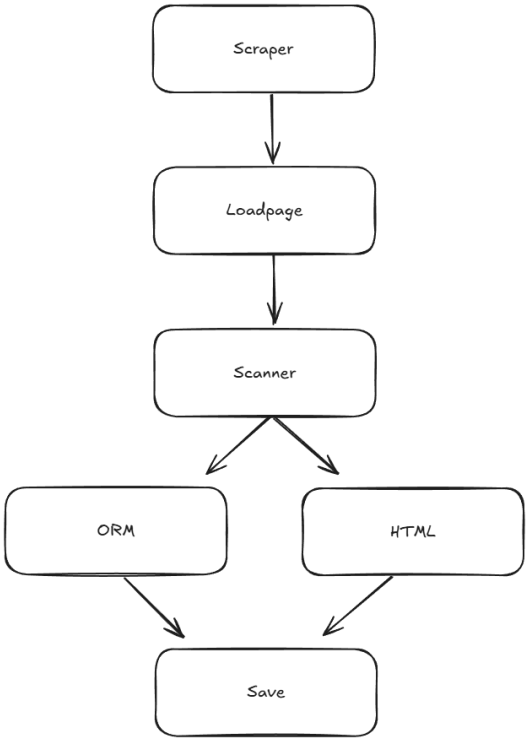
5.3.2.1 BETQUEUE

A funcionalidade BetQueue adiciona todas as bets ao sistema de filas de processamento, desde que não estejam previamente armazenadas no Redis. Este último gerencia a execução das tarefas agendadas, garantindo que, ao atingir ou exceder o tempo programado, o sistema inicie o processamento do WebScraper. Esse processo é realizado com o uso da biblioteca Puppeteer, que emprega o Node.js para emular o comportamento de um navegador e executar as tarefas necessárias.

5.3.2.2 FASTIFY

O servidor HTTP é inicializado por meio do Fastify, permitindo o gerenciamento das funcionalidades do sistema por uma API (Interface de Programação de Aplicações). Essa API oferece um painel de controle acessível, possibilitando ações como a inicialização manual do scraper, bem como o gerenciamento de outras funcionalidades essenciais do sistema.

5.4 Estrutura e Processos de Coleta e Análise de Dados



O pipeline de coleta e análise de dados foi desenvolvido para garantir a extração eficiente e organizada de informações das plataformas monitoradas, utilizando técnicas avançadas de Web Scraping e processamento OCR.

5.4.1 CARREGAMENTO DA PÁGINA (LOADPAGE)

O carregamento inicial da página é uma etapa crítica para a execução bem-sucedida do Scraper. O sistema é configurado para aguardar até que todos os elementos da página estejam totalmente inicializados antes de avançar. Esse processo assegura que as informações estejam

completas e prontas para análise, evitando inconsistências causadas por carregamento parcial ou dinâmico de conteúdo.

5.4.2 MAPEAMENTO (SCANNER)

Após o carregamento da página, é realizada uma varredura completa do DOM (Modelo de Objeto de Documento). Todos os elementos são identificados e armazenados para processamento posterior. Essa abordagem permite que o sistema lide de forma estruturada com os dados, otimizando o uso de recursos e evitando redundâncias durante a análise.

5.4.3 RECONHECIMENTO ÓPTICO DE CARACTERES (OCR)

Nessa etapa, o sistema realiza a análise OCR dos elementos capturados. Imagens são filtradas por tags como IMG e SVG, sendo processadas individualmente. A lógica implementada utiliza uma estrutura de dados do tipo Set, que elimina duplicatas, garantindo que cada elemento seja processado apenas uma vez.

Os elementos que possuem dimensões superiores à viewport padrão de 1920x1080 são descartados, uma vez que não são considerados representativos para a análise realizada. Todas as imagens processadas durante o funcionamento da ferramenta são armazenadas, possibilitando a criação de um banco de imagens. Essa abordagem evita o reprocessamento de imagens previamente analisadas e salvas, otimizando o desempenho do sistema. Após o armazenamento, as imagens são submetidas ao modelo OCR (Reconhecimento Óptico de Caracteres) PaddleOCR, sendo executado utilizando a linguagem de programação Python. A integração dessa funcionalidade à aplicação principal é feita por meio de uma API desenvolvida com o framework Flask, permitindo uma comunicação eficiente e modular entre os componentes do sistema.

5.4.4 HTML

Após a extração textual dos elementos presentes no HTML, o sistema realiza a análise detalhada das variáveis coletadas para identificar palavras ou frases de interesse. Durante esse

processo, são avaliadas propriedades como: cor do texto e do fundo, contraste conforme o padrão W3C, proporção em relação à Viewport (área de renderização de 1920×1080), distância em relação ao topo, visibilidade e localização.

5.4.5 SALVAMENTO (SAVE)

Por fim, todas as informações processadas são persistidas no banco de dados MySQL utilizando o TypeORM. Essa etapa finaliza o ciclo, garantindo que os dados estejam organizados e disponíveis para consultas e análises futuras.

6. LIMITAÇÕES E DESAFIOS

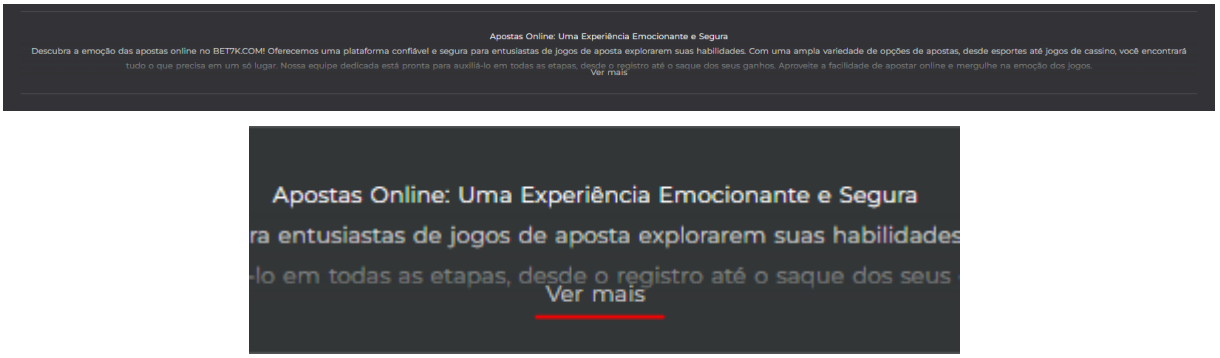
Durante o desenvolvimento e utilização do software, foram identificadas algumas limitações que podem impactar sua eficiência e funcionalidade. Essas restrições estão detalhadas a seguir.

A identificação de expressões-chave pode ser comprometida por variações no formato ou na apresentação dos textos. Diferenças na formatação ou desalinhamentos em imagens, como também o uso múltiplas tags HTML para compor elementos textuais, dificultam a detecção correta. Nesse contexto, é fundamental que o órgão regulamentador estabeleça padrões claros para o uso de elementos de texto e imagem. As imagens, por exemplo, devem incluir a propriedade “alt”, amplamente reconhecida como essencial para a acessibilidade, sendo utilizada para fornecer uma descrição ou transcrição fiel do conteúdo da imagem. Além disso, textos não devem ser compostos por múltiplas tags, especialmente em elementos que contenham cláusulas relacionadas a avisos de risco ou restrições etárias, garantindo uniformidade e clareza na apresentação das informações.

Fonte: https://mrjack.bet/

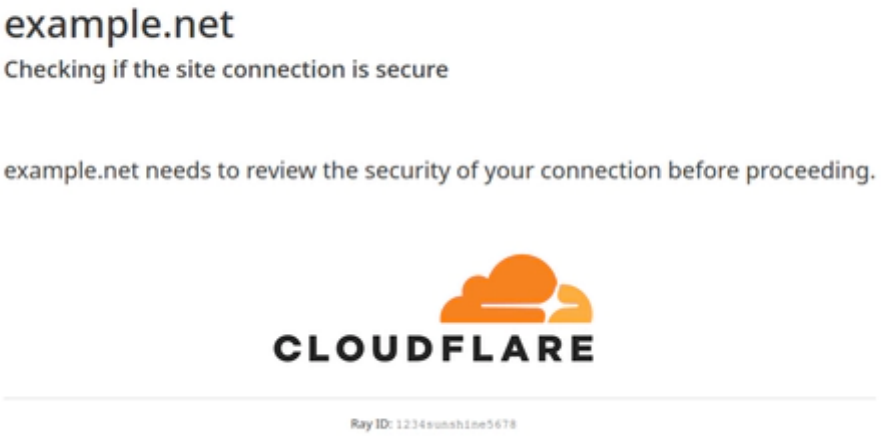
Nesse caso, O OCR não detecta ou interpreta corretamente o 18+. Pois o símbolo “+” não está devidamente alinhado com o texto “18”. Isso é uma limitação que necessitaria da criação de um modelo unicamente para detecção de elementos desse estilo, ou a criação como mencionado anteriormente de um padrão pré-estabelecido de como os elementos de imagem que contenham as cláusulas de restrição etária e aviso risco devem seguir.

É importante salientar que a ferramenta não realiza a análise de elementos que dependem de interação do usuário para serem carregados na página. Um exemplo comum são os casos em que há botões ou links, como "Mostrar mais", que acionam requisições a uma API interna do sítio de aposta para carregar conteúdo adicional. Esses elementos dinâmicos, por exigirem uma interação explícita do usuário, não são processados automaticamente pela ferramenta, limitando sua análise ao conteúdo disponível no carregamento inicial da página.



Fonte: https://bet7k.com/

Outro obstáculo importante é a presença de medidas anti-scraping, como CAPTCHAs e bloqueios de IP, implementadas por sítios eletrônicos de apostas para dificultar o monitoramento automatizado. Essas barreiras tornam necessárias soluções como uso de proxies ou atrasos entre as requisições, aumentando a complexidade do sistema e pode elevar os custos operacionais, como o uso de múltiplas máquinas com diferentes IPs.



Fonte: https://www.ctrl.blog/entry/cloudflare-ip-blockade.html

Além disso, o sistema de Reconhecimento Óptico de Caracteres (OCR) também enfrenta limitações, como dificuldade em interpretar textos em imagens de baixa qualidade, uso de fontes não padronizadas ou interferências gráficas. Apesar de técnicas como ajuste de contraste e remoção de ruídos serem aplicadas, a precisão não é garantida em todos os casos.

Fonte: https://onlybets.tv/promotions/regulation

Nesse caso, as cláusulas de advertência sobre risco e restrição etária ocupam áreas muito pequenas na imagem, tornando o processo de OCR ainda mais desafiador e complexo.

A eficácia da aplicação está diretamente relacionada a um ambiente configurado adequadamente, envolvendo a disponibilidade de servidores com capacidade computacional suficiente, instalação de pacotes específicos para a infraestrutura, como drivers NVIDIA compatíveis com os recursos computacionais, compatibilidade entre bibliotecas Python. Além disso, é essencial que os bancos de dados sejam devidamente dimensionados. Eventuais falhas

ou inadequações na infraestrutura podem comprometer o desempenho do sistema, impactando negativamente sua escalabilidade e confiabilidade.

Portanto, é essencial que o sistema continue a ser aprimorado, visando superar esses desafios e garantir maior eficiência e confiabilidade em sua utilização.

7. CONSIDERAÇÕES FINAIS

O desenvolvimento desta aplicação representou um importante passo para o monitoramento contínuo dos sítios eletrônicos de apostas no Brasil. A integração de tecnologias modernas, como o Fastify, Puppeteer, OCR com PaddleOCR, e o gerenciamento de tarefas com Redis e Bull, permitiu a construção de um sistema funcional, escalável e eficiente. Os resultados obtidos até o momento demonstram que a aplicação tem o poder de identificar possíveis irregularidades nas plataformas, contribuindo para a conformidade com as normas legais e promovendo práticas de jogo responsável.

7.1 APRIMORAMENTOS FUTUROS

Uma das melhorias previstas é a criação de uma interface de usuário intuitiva, que facilite a interação com a ferramenta, tornando-a mais acessível e eficiente. Além disso, o modelo de Reconhecimento Óptico de Caracteres (OCR) será otimizado para aprimorar a precisão na extração de textos a partir de imagens, o que proporcionará maior confiabilidade nos resultados obtidos.

Outro aprimoramento envolve o desenvolvimento de um sistema personalizado capaz de identificar conteúdos gráficos que sejam restritos a maiores de idade ou que contenham avisos de risco. Isso garantirá um maior controle sobre o cumprimento das normas regulatórias.

Por fim, a implementação da funcionalidade de execução distribuída permitirá que a ferramenta seja executada simultaneamente em várias máquinas, ampliando sua escalabilidade e eficiência no processamento de dados.

REFERÊNCIAS

Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseract: *A case study. International journal of computer applications*, 55(10), 50-56. Disponível em: https://www.academia.edu/download/100190454/3e47cc647c47a1a249e1103047dd5b002b5a.pdf. Acesso em: 15 nov. 2024.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3). Disponível em: http://www.i-csrs.org/Volumes/ijasca/2021.3.11.pdf. Acesso em: 15 nov. 2024.

Keller, M. S. (1999). Take command: cron: Job scheduler. Linux Journal, 1999(65es), 15-es. Disponível em: https://dl.acm.org/doi/fullHtml/10.5555/327966.327981. Acesso em: 16 nov. 2024.

Indra, J., & Sarjono, H. (2010). Queue Analysis System For Improving Efficiency Of Service. Jurnal Manajemen Teori dan Terapan *Journal of Theory and Applied Management*. Disponível em: https://pdfs.semanticscholar.org/55ce/cc3fe41aa3c2c2d152c4cfe799115adc6b0a.pdf. Acesso em: 16 nov. 2024.

ANEXO X CONAR Disponível em: http://www.conar.org.br/pdf/CONAR-ANEXO-X-PUBLICIDADE-APOSTAS-dezembro-2023.pdf. Acesso em: 17 nov. 2024.