Sistema Orientado por PLN para Classificação de Discurso Ofensivo em Ambientes Escolares

NPL-Driven System for Offensive Speech Classification in School Environmentes

Bruno Freitas {bruno.oliveira249@fatec.sp.gov.br}
Caio Moraes {caio.moraes@fatec.sp.gov.br}
Daniel Mandira {daniel.mandira@fatec.sp.gov.br}
Isabela Chaves {isabela.pedroso@fatec.sp.gov.br}
Mauricio Bertoldo {mauricio.oliveira23@fatec.sp.gov.br}
Ruth Mendonça {ruth.mendonca@fatec.sp.gov.br}

RESUMO

O projeto Resist tem como objetivo criar uma solução robusta para identificar e restringir conteúdos com contextos de discursos injuriosos acessados na internet dentro de instituições escolares, garantindo um ambiente digital mais seguro e controlado. A proposta combina tecnologias avançadas, como o servidor proxy Squid, utilizado para monitoramento e filtragem de páginas, e inteligência artificial baseada na arquitetura BERT, uma ferramenta poderosa para compreensão semântica de textos.

A iniciativa visa não apenas bloquear conteúdos tóxicos, como discursos de ódio, mas também aprimorar continuamente sua capacidade de detecção por meio de aprendizado contínuo. Dados registrados de acessos são processados e analisados em um banco de dados NoSQL (MongoDB), permitindo que a IA classifique os conteúdos em categorias como insultos, ameaças, e ataques à identidade. Caso um conteúdo seja identificado como nocivo, ele é bloqueado e utilizado para reforçar o treinamento do modelo, aumentando sua eficiência e precisão ao longo do tempo.

Além de oferecer proteção em tempo real, o sistema reflete um compromisso com a inovação tecnológica e a responsabilidade digital, buscando proteger usuários e organizações contra interações prejudiciais online.

PALAVRAS-CHAVE: Discriminação; Inteligência Artificial; Discurso de ódio; BERT; PLN.

ABSTRACT

The Resist project aims to create a robust solution to identify and restrict harmful content with injurious discourse contexts accessed on the internet within educational institutions, ensuring a safer and more controlled digital environment. The proposal combines advanced technologies, such as the Squid proxy server, used for monitoring and filtering web pages, and artificial intelligence based on the BERT architecture, a powerful tool for semantic text comprehension.

The initiative seeks not only to block toxic content, such as hate speech, but also to continuously improve its detection capability through ongoing learning. Access logs are processed and analyzed in a NoSQL database (MongoDB), enabling the AI to classify content into categories such as insults, threats, and identity attacks. If a piece of content is identified as harmful, it is blocked and used to retrain the model, enhancing its efficiency and accuracy over time.

In addition to providing real-time protection, the system reflects a commitment to technological innovation and digital responsibility, aiming to safeguard users and organizations against harmful online interactions.

KEYWORDS: Discrimination; Artificial Intelligence; Hate speech; BERT; NLP.

INTRODUÇÃO

A Organização das Nações Unidas (ONU), fundada em 1945, tem como objetivo promover a cooperação internacional entre os países, focando-se nos direitos humanos, desenvolvimento econômico e segurança mundial. Durante as últimas décadas, a ONU expandiu seus horizontes para problemas atuais, impulsionada pela necessidade de abordar assuntos como a desigualdade social e mudanças para garantir uma educação mundial de qualidade, de forma que foi criada a Agenda 2030, que consiste em 17 Objetivos de Desenvolvimento Sustentável (ODS).

Este artigo busca relacionar-se aos objetivos da organização, aliando-se especificamente a meta quatro, que visa assegurar educação inclusiva e equitativa de qualidade, assim como a meta dez, que promove a inclusão social, econômica e política de todos independentemente de idade, sexo, raça, etnia, origem, religião, condição social e econômica etc. Ambas as ODS traçam um plano universal para alcançar um futuro melhor, especificamente no âmbito social e educacional.

Trazendo o contexto para o discurso de ódio na internet, sua incidência tem aumentado significativamente, acompanhando o crescimento do uso das redes sociais, que são utilizadas por aproximadamente 60% da população mundial (Ltd, 2020). A internet, como espaço de troca constante de ideias, proporciona liberdade de expressão, mas o anonimato que a acompanha tem sido, em muitos casos, mal utilizado. Essa liberdade tem permitido que indivíduos expressem suas opiniões de forma desmedida, levando, por vezes, a discursos ofensivos ou comportamentos tóxicos. A ameaça constante de assédio ou comentários nocivos na web tornou-se um obstáculo para compartilhar ideias e interesses de forma segura. Essa toxicidade online, que frequentemente resulta em comportamentos de humilhação, perseguição e discriminação, é amplamente conhecida como Cyberbullying.

Na União Europeia, por exemplo, 80% das pessoas com acesso à internet relataram ter presenciado discursos de ódio online, enquanto 40% afirmaram ter se sentido pessoalmente atacadas ou ameaçadas nesses espaços virtuais (Castaño-Pulgarín *et al.*, 2021). Já na Coreia do Sul, uma pesquisa realizada pela "National Human Rights Association of Korea"em 2019, envolvendo 1.200 adultos e 500 jovens, revelou dados alarmantes. No grupo de adultos, 64% relataram ter sido expostos a diferentes formas de discurso de ódio no ano anterior. Entre os principais tipos, destacaram-se ataques relacionados ao local de nascimento (74,6%), sexismo contra mulheres (68,7%), discursos contra pessoas idosas (67,8%), preconceito contra minorias sexuais (67,7%), imigrantes (66%) e pessoas com deficiência (58,2%). No grupo jovem, 68,3% indicaram ter vivenciado exposições a discursos de ódio, com os casos mais comuns sendo preconceito contra mulheres (63%) e minorias sexuais (57%). Importante ressaltar que 83% das ocorrências foram registradas em redes sociais e outros ambientes digitais (Lee; Kim; Kim, 2019).

No Brasil, o discurso de ódio está amplamente presente tanto nas redes sociais quanto em espaços digitais e físicos, não se limitando ao ambiente online. Em uma análise de 145 notícias publicadas em um único dia (7 de junho de 2016), constatou-se que 90% delas continham ao menos um comentário de ódio (Pelle; Moreira, 2016). Em muitos desses casos, usuários engajados em discussões sobre as notícias acabavam se envolvendo em conflitos, promovendo ataques e ofensas. Outro estudo relevante foi conduzido pela agência Nova/SB, que investigou a intolerância em redes sociais ao longo de três meses, monitorando plataformas como Facebook, Twitter e Instagram. Nesse período, foram registrados 542.781 comentários com teor de discurso de ódio, sendo 84% deles relacionados a questões negativas, como racismo, misoginia e xenofobia. Esses dados reforçam a necessidade de iniciativas que combatam ativamente o discurso de ódio e promovam a conscientização sobre o impacto dessas práticas nos ambientes digitais e sociais.

Os desafios no combate ao Cyberbullying incluem a identificação ineficiente da toxicidade extrema e a distinção entre "predadores"e vítimas nesse ambiente digital. Redes sociais como Facebook e X (antigo Twitter), onde essas práticas são mais comuns, já incorporam sistemas automatizados para identificar e mitigar agressões e assédios em suas plataformas. Contudo, tais ferramentas frequentemente enfrentam limitações de precisão devido à falta de refinamento nas metodologias de busca e análise do conteúdo (Sharma *et al.*, 2017).

Outro ponto relevante, é de que a fase de formação da criança e é de extrema relevância para a construção dos valores da criança em desenvolvimento como também de uma fases de transição do seu eu, e de suas relação sociais. Assim como, as escolas podem criar um ambiente que venha a

constituir-se num "espelho"e num "mundo"para as crianças, ajudando-as a caminhar para fora de um ambiente familiar adverso (Szymanski, 2007).

Portanto, estes dados e desafios evidenciam a necessidade de soluções mais avançadas, que consigam detectar nuances e contextos complexos de discurso de ódio na web, abrangendo categorias como sexismo, homofobia, xenofobia, racismo etc., tornando possível a promoção de um ambiente digital mais seguro e inclusivo, principalmente em instituições de ensino, para que dessa forma, elas sejam capazes de conscientizar alunos que acessam ou apresentam falas de cunho preconceituoso.

Diante disso, a área de Tecnologia da Informação (TI) se destaca como um campo estratégico para o desenvolvimento de soluções inovadoras capazes de abordar questões complexas, como o discurso de ódio na internet. Com os constantes avanços na Inteligência Artificial (IA), surgem oportunidades promissoras para automatizar processos de auditoria de conteúdo web, especialmente no que diz respeito à identificação e mitigação de discursos de ódio. A IA desponta como uma ferramenta poderosa nesse cenário, oferecendo a capacidade de analisar grandes volumes de dados em tempo real por meio de algoritmos avançados e técnicas de aprendizado de máquina. A detecção de conteúdo injurioso, não apenas nas redes sociais, mas em todo o ambiente web — particularmente no contexto escolar —, é uma medida essencial para garantir um ambiente de convivência pacífica.

Este projeto, portanto, desenvolve uma solução que permite identificar e bloquear automaticamente esses contextos acessados no ambiente escolar, utilizando dados coletados em situações reais para atingir resultados robustos e confiáveis, assim como empregando o Processamento de Linguagem Natural (PLN) para identificar e bloquear sites que apresentem contextos de discurso de ódio.

O PLN, uma subárea da IA, concentra-se na interação entre computadores e a linguagem humana. Ele utiliza algoritmos e técnicas de análise textual que permitem aos sistemas computacionais compreender, interpretar e gerar linguagem humana de forma semelhante aos seres humanos. No contexto deste projeto, o PLN será empregado para analisar o conteúdo textual de sites acessados por alunos, identificando padrões linguísticos associados a discursos de ódio. Todos os dados utilizados para o treinamento da IA foram ajustados por meio do modelo pré-treinado Bidirectional Encoder Representations from Transformers (BERT).

O BERT, desenvolvido pelo Google, é um modelo de aprendizagem profunda projetado especificamente para tarefas de Processamento de Linguagem Natural. Ele é reconhecido por sua capacidade de compreender o contexto completo das palavras dentro de uma frase, graças ao seu treinamento bidirecional. Diferentemente de outros modelos, como o Word2Vec ou o GloVe, o BERT analisa as palavras considerando todo o contexto da sentença, o que o torna mais eficaz na identificação de nuances linguísticas complexas (Saleh; Alhothali; Moria, 2023).

A implementação de um sistema de auditoria de conteúdo web baseado em PLN oferece inúmeros benefícios para instituições de ensino. A automação proporcionada por todo o sistema reduz significativamente o esforço manual necessário para auditorias, tornando o processo mais eficiente e escalável. Em última análise, a adoção desse sistema reafirma o compromisso das escolas com a promoção dos direitos humanos, igualdade e respeito à diversidade, formando cidadãos mais conscientes e engajados na construção de uma sociedade justa e inclusiva.

OBJETIVO

a) Desenvolver um algoritmo de PLN capaz de analisar o conteúdo textual de páginas da *Web* e identificar padrões linguísticos associados a conteúdos discriminatórios, injuriosos ou que incitem discurso de ódio.

b) Integrar o sistema de auditoria de conteúdo *Web* com um sistema de notificação, de forma

que o usuário seja alertado na tela quando for identificado conteúdo ofensivo ou prejudicial, incluindo informações sobre o motivo do bloqueio e opções para revisão por administradores escolares, caso necessário.

c) Funcionalidade de retroalimentação, permitindo que o sistema aprenda continuamente com novos dados, aprimorando sua capacidade de identificação e bloqueio de conteúdos prejudiciais através do feedback dos administradores ou de atualizações periódicas de dados e padrões linguísticos.

d) Desenvolver um painel de controle administrativo, que permita a visualização de estatísticas sobre tentativas de acesso bloqueadas, tipos de conteúdo identificado, e histórico de notificações, oferecendo também a possibilidade de personalizar critérios de bloqueio de acordo com as políticas da instituição.

ESTADO DA ARTE

O estudo realizado em Pitropakis *et al.* (2020), teve como foco o combate ao discurso de ódio contra imigrantes, por ser um grande problema na América do norte e em regiões Europeias. Foi desenvolvido um programa que utiliza PLN para identificar postagens xenofóbicas na plataforma X (anteriormente chamada Twitter), por ser muito utilizada para escrever e compartilhar pequenos textos opinativos, característica que abre espaço para muitas postagens preconceituosas e agressivas. A composição do conjunto de dados foi feita utilizando a API pública do Twitter, que disponibiliza dados públicos gerados por usuários, por onde foram coletadas diversas postagens feitas em inglês nos EUA, Canadá e Reino Unido, que continham mensagem anti-imigrante e/ou xenofóbica. Foram obtidas 8270 postagens, manualmente separadas em negativas, não negativas, indeciso e não relacionado.

Após o tratamento dos dados, optou-se pela utilização de alguns algoritmos comumente empregados em classificação de texto, como Support Vector Machines (SVM), Naïve–Bayes (NB), e Logistic Regression (LR), sendo os três testados para escolher o mais adequado. Estes performam bem se empregados na detecção de postagens com discursos de ódio, detecção de linguagem abusiva. Além da capacidade de encontrar as palavras mais frequentes entre os tweets, o modelo obteve uma precisão de 87% na classificação de conteúdo xenofóbico, com a técnica LR aplicada a N-gramas com palavras. N-gramas são sequências de palavras, portanto, um algoritmo que trabalhe com elas poderá calcular a probabilidade da palavra seguinte em uma frase, a partir de um conjunto de dados com diversas frases. É a mesma metodologia utilizada nos *browsers*, para autocompletar pesquisas antes que o usuário termine de digitar Srinidhi (2019). No caso de Pitropakis *et al.* (2020), as frases com probabilidades semelhantes indicam contextos semelhantes, possibilitando a classificação das frases.

Dos desafios encontrados no desenvolvimento, destacam-se a dificuldade na definição do que é xenofobia, pois trata-se de um termo muito amplo, sendo um tipo de preconceito que pode ser direcionado a diversos povos diferentes. A partir disso, surgem inúmeros xingamentos e injúrias relacionados à cultura ou peculiaridades de um povo. Portanto, o foco foi em termos mais abrangentes, como imigrante, imigração e refugiados, o que, certamente deixou de fora muitas postagens mal-intencionadas. Houve dificuldade também em capturar o sentimento negativo referente aos imigrantes, pois em uma postagem, ele pode se apresentar de diversas maneiras, sendo mais aparentes com o uso de palavrões, ou mais discretas, como é o caso de comentários sarcásticos, que necessitariam de um treinamento específico para reconhecê-los. Em sua metodologia, este monitoramento de comentários anti-imigrantes assemelha-se muito à proposta deste projeto, portanto, estas mesmas dificuldades são consideradas no desenvolvimento do sistema ReSist.

Em Pelle Pelle e Moreira (2017), a verificação é ainda mais ampla, pois o projeto tem como objetivo a detecção de comentários ofensivos na *Web* brasileira. Além das dificuldades de se trabalhar

com um conceito tão abrangente, o autor também pontua que muitos usuários disfarçam palavras ofensivas trocando letras por números ou símbolos, portanto, a simples criação de uma lista de bloqueio poderia deixar de fora muitos comentários, daí surge a necessidade de um conjunto de dados com exemplos positivos e negativos. Este conjunto foi criado utilizando como base o site de notícias G1, que por ser o mais acessado do país, possui muitos comentários em suas notícias. Verificou-se que em cerca de 90% das notícias analisadas, havia pelo menos um comentário ofensivo. Com a implementação de um *Webscraper*, um software para coletar dados de sites, foram extraídas 115 notícias, com 10,366 comentários. Destes, foram selecionados 1,250 comentários, a serem classificados como ofensivos ou não, e em caso afirmativo, classificados como racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou palavrões. Para a classificação foi desenvolvida uma ferramenta *Web* que exibe o comentário, o link da notícia e as opções de classificação, o que agilizou todo o processo.

A seguir, são tomadas algumas medidas de pré-processamento, como o *case folding*, que consiste em converter todas as letras para o mesmo tipo de caixa, no caso, caixa baixa, e a conversão em n-gramas, usando unigramas, bigramas e trigramas (1 palavra ou sequências de 2 e 3, respectivamente), para capturar a estrutura do comentário odioso. Foi testada também a técnica de seleção de *features*, que é o processo de redução do número de variáveis de entrada no desenvolvimento de um modelo preditivo. Buscando remover variáveis que podem retardar o desenvolvimento e treinamento do modelo, reduzindo custo computacional e melhorando desempenho, em alguns casos. Pelle Pelle e Moreira (2017) também utiliza os Algoritmos de Naive Bayes e SVM.

O modelo foi avaliado a partir do F-score, calculado a partir da medida de resultados falsos-positivos (precisão, ou PPV), e falsos-negativos(revocação, ou *recall*) Dantas, Marcel (2019). O melhor desempenho obtido foi de .85, utilizando SVM. Os pesquisadores comentam que o contexto é essencial para a classificação, pois em diversos casos, as mesmas palavras que compõem um comentário ofensivo, não seriam ofensivas em outra frases.Foram testados os bigramas e trigramas com o objetivo de realizar esta verificação de forma adequada e obter melhor performance de classificação, no entanto, verificou-se que o uso de n-gramas mais longos não necessariamente aumenta a precisão do modelo, visto que os resultados foram semelhantes com os 3 tipos. Sendo assim, torna-se preferível a utilização de unigramas, por custarem menos poder de processamento. Observa-se também que a linguagem na Web está sempre repleta de jargões novos, abreviações e erros de gramática, logo, um modelo a ser aplicado neste ambiente deve ser capaz de se adaptar a estas e outras mudanças.

Sobre o discurso de ódio, Poojitha *et al.* (2023) emprega o uso de Processamento de Linguagem Natural (PLN), fazendo a classificação dos textos, separando-os por índice de toxicidade. Durante o processo de classificação foram comparados diversos modelos diferentes, incluindo o BERT (Bidirectional Encoder Representations from Transformers), XLNet e CNN (Convolutional Neural Network), foi identificada certa superioridade no modelo BERT, empregado agora como o modelo a ser usado definitivamente, por ter se saído melhor na classificação de toxicidade dentro dos textos de forma automática.

Para testes foram utilizados diferentes tipos de técnicas dentro de cada modelo. Um deles sendo a BoW (Bag of Words), transformando partes do texto em "sacos", ou seja, conjuntos de palavras sem nexo umas com as outras, e depois conta quantas vezes cada palavra aparece, criando uma matriz onde cada palavra equivale a uma característica, e esta matriz representa a frequência de cada palavra no texto.

De maneira resumida, este projeto tem o objetivo de fazer testes práticos com diferentes modelos de aprendizagem de máquina, a fim de determinar qual teria maior desempenho final e maior

precisão na detecção destes comentários. Utilizando de um dataset que possui cerca de 158.640 comentários, durante este projeto, foram extraídos comentários tóxicos e não-tóxicos, fazendo o pré-processamento dos dados em seis etapas, como a *Tokenização*, que é o processo de dividir o texto em partes menores chamadas Tokens, que podem ser números, palavras ou quaisquer símbolos que contenham informação. A segunda etapa consiste na remoção da pontuação e números destes comentários (de forma automática, utilizando PLN), o que torna mais rápido o processamento e aprendizado da Inteligência Artificial (IA). O quarto passo seria chamado de *Stemming*, que consiste em transformar uma palavra para sua forma mais "crua e básica", como por exemplo a palavra "jogando", que é uma forma alterada da palavra "jogo", e é algo que pode ser implementado fazendo uso de algoritmos. O quinto passo é a correção de palavras com erros gramaticais, substituindo-as pelas palavras corretas. O sexto e último passo é a remoção de certas palavras (neste projeto são chamadas de *Stopwords*, termo em inglês que se refere a palavras que não alteram o sentido geral da frase), com o intuito de diminuir a complexidade destas frases e otimizar o aprendizado da IA.

O estudo possui a finalidade de avaliar o desempenho de diferentes modelos de aprendizado de máquina para a classificação de conteúdo tóxico envolvendo comentários pré-coletados em redes sociais. No presente contexto, depois de fazer o pré-processamento dos dados, a rede neural identifica dentro do dataset todos os comentários tóxicos e não-tóxicos, os categorizando por estes tipos.

Dos artigos correlatos, este emprega mais similaridade tanto em seu conteúdo-chave (textos com conteúdo de ódio), quanto em sua empregabilidade e processos. A utilização do modelo BERT se tornará indispensável para o andamento do projeto ReSist, assim como o uso de PLN e aprendizagem de máquina na classificação dos textos.

METODOLOGIA

Neste projeto, será desenvolvido um sistema de auditoria de conteúdo *Web*, utilizando técnicas de PLN, que proporcione a identificacão e o bloqueio de sites acessados no âmbito escolar que contenham contextos discriminatórios. Utilizando IA, por meio do Processamento de Linguagem Natural (PLN). Quando um aluno acessar um site na *Web*, o programa, escaneará todo o conteúdo da página em busca de contexto de discriminação. Caso haja algo, o programa atuará por meio de um proxy, para bloquear o site em toda a rede. Caso contrário, o site será aberto normalmente no navegador. O proxy é um aplicativo do servidor que atua na intermediação entre um cliente que solicita um recurso e um servidor que o fornece, portanto, ele é capaz de monitorar e controlar o acesso a sites em uma rede.

Para as primeiras implementações do sistema de bloqueio automatizado, a tecnologia utilizada no o back-end será Python, escolhida pela sua vasta gama de bibliotecas que auxiliarão no tratamento e manipulação do conteúdo das páginas. O Ambiente de Desenvolvimento Integrado (IDE) escolhido para manipular a linguagem foi o Visual Studio Code, notável pela sua flexibilidade e suporte de diversas extensões, o que auxiliam no processo de codificação.

O fluxo do sistema respresentado no Fluxograma 1 se inicia quando o cliente faz uma solicitação ao servidor/proxy (Etapa A). Para atuar como servidor, utilizaremos o Ubuntu, uma das distribuições mais populares do sistema operacional Linux. É conhecido por sua facilidade de uso, estabilidade e segurança.Canonical Ltd. (s.d.) Nele, será instalado um *Software* de Proxy conhecido como Squid, que permite monitorar as páginas acessadas pelos computadores conectados, bem como impede o acesso a determinados *websites*.(The Squid Software Foundation, s.d.), e atua como intermediário entre o cliente e a internet (Etapa B). Em seguida, a solicitação do cliente é enviada à internet para buscar o conteúdo requisitado (Etapa C). A cada solicitação, o sistema registra informações de acesso, como data, hora, IP da máquina e URL do site acessado, e armazena esses

dados no arquivo access.log do servidor (Etapa D).

Na etapa seguinte, o módulo indexador registra no banco de dados as informações do acesso mais recente registrado no access.log (Etapa E). No projeto será implementado o modelo NoSQL. A escolha pelo MongoDB se deve pela sua orientação de documentos do tipo JavaScript Object Notation (JSON), que é um formato leve e fácil de usar para armazenar e transportar dados. O MongoDB possui capacidade de lidar com grandes volumes de dados não estruturados de maneira eficaz, além de sua alta disponibilidade e suporte à replicação, fornecendo garantia a continuidade do sistema em caso de falhas. A finalidade dessa etapa é aprimorar a capacidade de monitoramento do sistema. Os dados registrados incluem o caminho do arquivo com o conteúdo indexado (pathLocal), um indicador do tipo booleano que sinaliza se o acesso está bloqueado ou não (flag), a URL acessada (urlWeb) e o termo identificado que pode ter causado o bloqueio.

Posteriormente, o conteúdo é armazenado no banco de dados, e enviado para avaliação e classificação pelo Módulo de Inteligência Artificial (Etapa F). Nesse momento, a Inteligência Artificial (IA) identifica e classifica o conteúdo com base nos parâmetros estabelecidos (Etapa G). A IA desenvolvida neste projeto tem como foco principal identificar contextos discriminatórios em conteúdos acessados na internet, utilizando técnicas avançadas de PLN. O modelo foi construído com base na arquitetura BERT, uma tecnologia de aprendizado profunda desenvolvida pelo Google que permite a compreensão do contexto das palavras em um texto de forma bidirecional. Isso significa que o modelo analisa uma palavra considerando tanto o que vem antes quanto o que vem depois, garantindo maior precisão na interpretação semântica.

O treinamento da IA foi realizado utilizando o *dataset* público *Jigsaw Unintended Bias*, amplamente reconhecido em estudos relacionados à detecção de discurso de ódio. Este *dataset* contém milhões de exemplos rotulados de comentários online e textos, além de que, ele inclui diversas categorias de discurso ofensivo, como toxicidade geral, insultos, ameaças, ataques à identidade, conteúdo obsceno e sexual explícito. Antes do treinamento, os dados passaram por etapas de pré-processamento, como limpeza textual e tokenização, que converte os textos em entradas numéricas compatíveis com o modelo BERT, permitindo que ele processe e aprenda padrões relevantes para a tarefa.

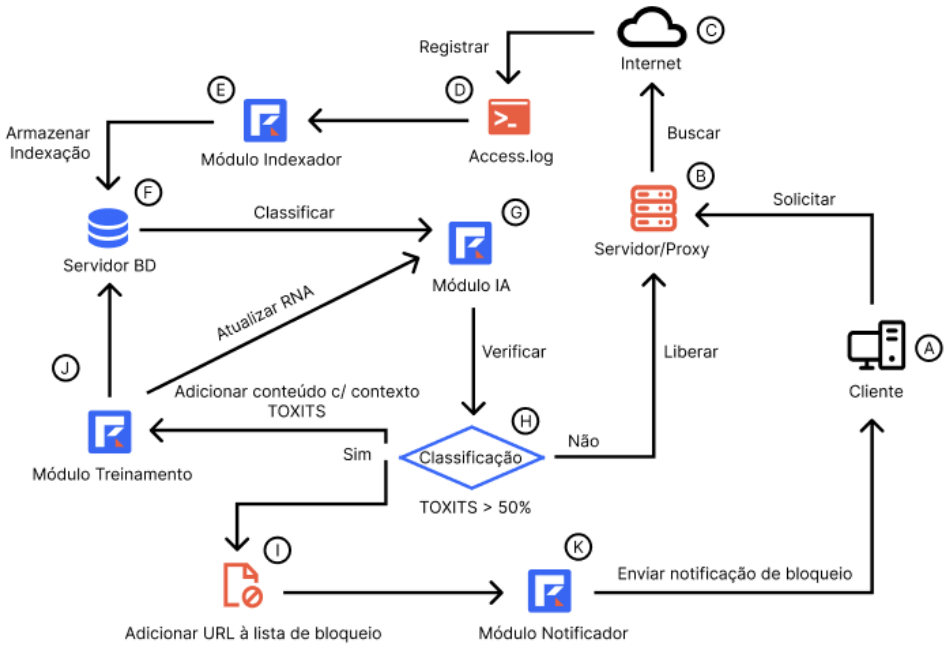
Com os resultados dessa análise, o sistema verifica o nível de toxidade do texto e o classifica em: Toxicity (tóxico), Severe Toxicity (severamente tóxico), Obscene (obsceno), Sexual Explicit (sexualmente explicito), Identity Attack (ataque de identidade), Threat (ameaça) e Insult (insulto). Caso o nível de toxidade (TOXITS) seja inferior a 50

Se o conteúdo for considerado tóxico, ou seja, com classificação superior a 50%, ele é adicionado a uma lista de bloqueio no servidor denominada bloqueados.txt, impedindo acessos futuros ao mesmo conteúdo (Etapa I).

Além disso, após a classificação feita pelo Módulo IA, se o conteúdo for considerado tóxico, ele é encaminhado ao Módulo de Treinamento para aprimorar a IA. O conteúdo tóxico é utilizado no treinamento da IA sendo armazenado no banco de dados com informações de contexto, servindo como referência para futuras classificações. A Rede Neural Artificial (RNA) da IA é então atualizada com esses novos dados, permitindo a retroalimentação do sistema e garantindo que ele continue a evoluir e se adaptar, melhorando a eficiência na detecção e classificação de conteúdos potencialmente prejudiciais (Etapa J).

Por fim, quando uma URL é adicionada à lista de bloqueio ou o cliente solicita acesso a uma URL já bloqueada, o sistema exibe uma notificação informando que o acesso foi restrito devido ao alto nível de TOXITS (Etapa I).

Fluxograma 1 – Resist



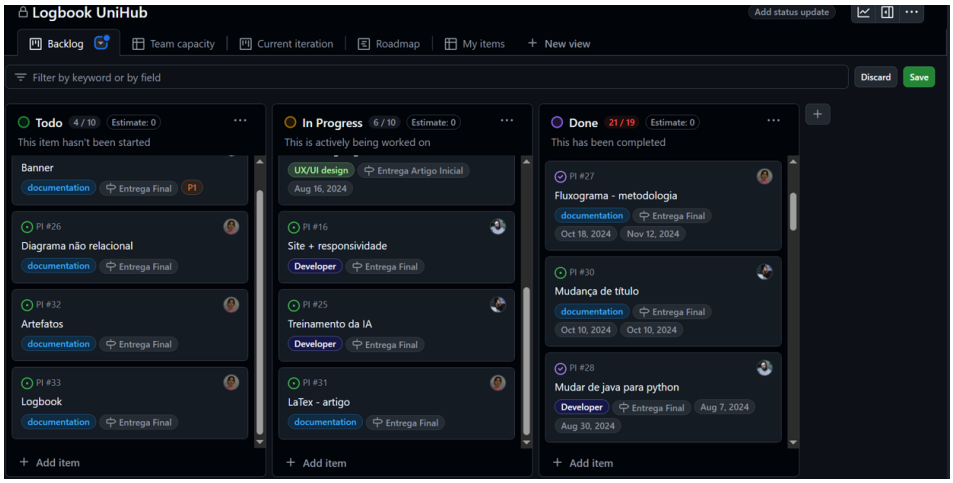
Fonte: Autoria Própria (2024)

RESULTADOS PRELIMINARES

KANBAN

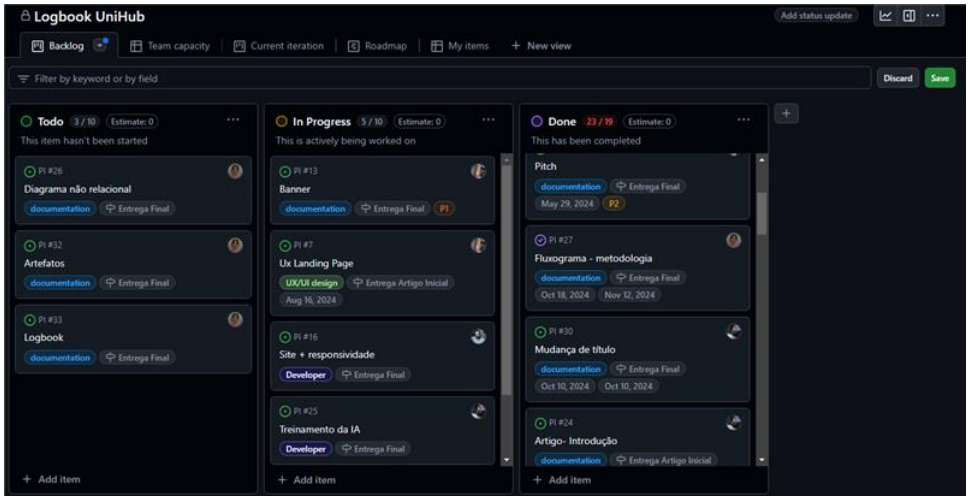
O monitoramento pelo logbook do GitHub, torna a organização do projeto ainda mais eficaz. O GitHub permite um acompanhamento detalhado de cada etapa, com transparência total nas tarefas e progresso da equipe. O logbook facilita o registro de cada atividade. Isso resulta em uma gestão de projeto ágil, organizada e colaborativa, garantindo a entrega de valor em ciclos curtos e controlados. A seguir três fases do kanban durante uma semana: Na Figura 1, há quatro tarefas para começar; já na Figura 2, a tarefa "Banner"está em progresso; na Figura 3 já não há mais tarefas para serem iniciadas, todas foram finalizadas apenas "LaTex - artigo"esta em progresso.

Figura 1 – 1º visão



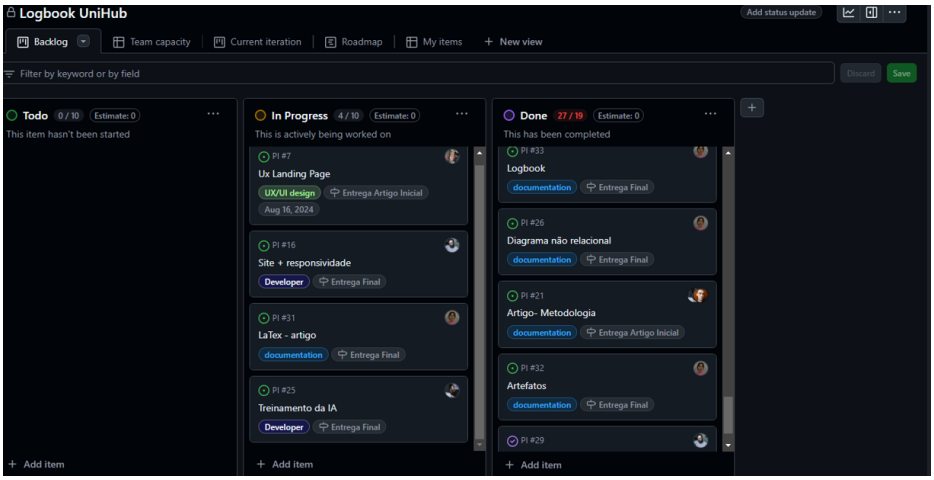
Fonte: Autoria Própria (2024)

Figura 2 – 2º visão



Fonte: Autoria Própria (2024)

Figura 3 – 3º visão

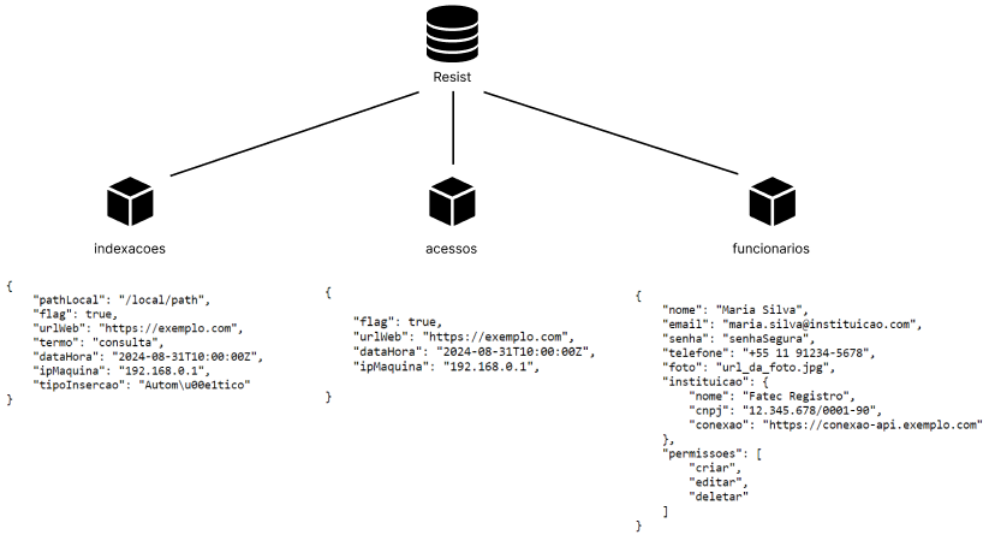


Fonte: Autoria Própria (2024)

DIAGRAMA DE BANCO DE DADOS NÃO RELACIONAL

O diagrama foi pensando na seguinte forma, a tabela indexações registra os dados relacionados ao conteúdo do site em si, como local de armazenamento, url da página, se será liberada ou bloqueada, e o termo que justifica este possível bloqueio. Além das informações de data e máquina que ocasionou essa indexação, por acessá-la. Já a tabela acessos, registra os dados dos demais acessos a um site já verificado e indexado, por isso apenas possui. Esta tabela serve para fins estatísticos, portanto, apresenta a url da página, data, hora e ip da máquina por onde foi acessado. A tabela funcionarios registra os dados relacionados aos colaboradores de uma instituição. Cada documento armazena informações pessoais, como nome, e-mail, telefone e uma URL para foto de perfil do usuário. Além disso, inclui o campo instituicao, que é um documento aninhado, em qual são armazenados os dados de outras instituições que o usuário possa estar vinculado, como o nome, CNPJ e o *endpoint* da *Application Programming Interface* (API) utilizado para integrações ou consultas relacionadas à instituição. Também são registradas as permissões atribuídas a cada funcionário, detalhando os tipos de ações que podem ser realizadas no sistema, como criar, editar ou deletar informações.

Figura 4 – Diagrama de Banco de Dados NoSQL



Fonte: Autoria Própria (2024)

APLICAÇÃO DESKTOP

O desenvolvimento da aplicação Desktop foi realizada utilizando a linguagem de programação Python, este sistema desempenha o papel de extrair o conteúdo dos sites, para que possam ser analisados em busca de possíveis contextos injuriosos, e caso seja encontrado, bloqueá-los na rede local.

Utilizando o *Requests*[1], obteve-se um sistema capaz de monitorar os acessos relatados pelo Squid. A aplicação retorna para cada acesso registrado a Url, data, hora e ip da máquina que solicitou o acesso. Neste ponto, a aplicação confere se o site já foi verificado antes, e em caso negativo, realiza todo o processo de extração e verificação do teor da página.

No entanto, a aplicação ainda não é capaz de verificar contextos, pois ainda não houve a integração com a Inteligência Artificial, por limitações com o hardware usado no treinamento, o que limita a detecção do sistema. Portanto, o programa realiza uma leitura da página em busca de uma palavra específica, e a presença ou ausência desta palavra será o fator determinante para bloqueio ou permissão do acesso a esta página, caso determinado site seja bloqueado uma notificação é enviada ao usuário (Figura x) e a página não poderá ser vista pelo usuário (Figura Y).

A Url da página é armazenada em um banco de dados, para que seja possível visualizar se a página está liberada ou não, e qual palavra ou frase motivou seu bloqueio. Após a verificação, o sistema também utiliza os dados recuperados para registrar no banco cada acesso, o que permitirá a geração de dados detalhados, que serão utilizados para criar gráficos e relatórios acessíveis via Web para os gestores das instituições.

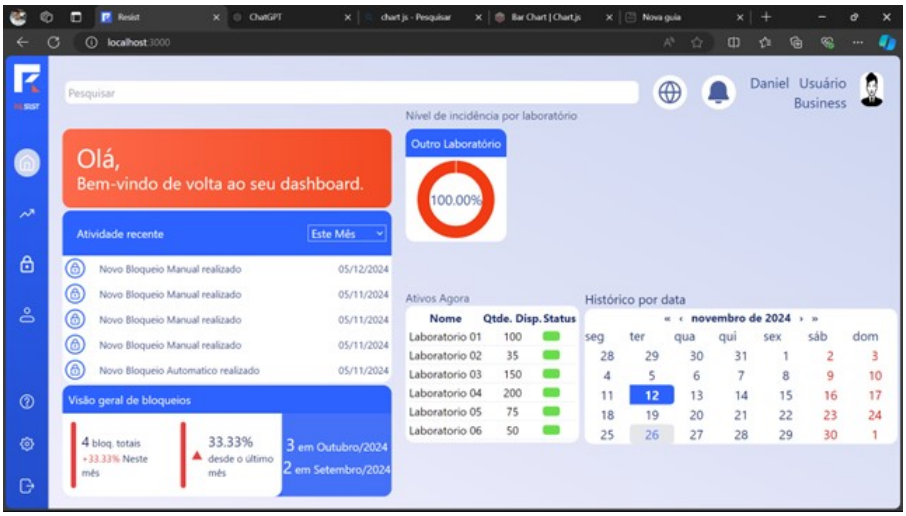
WEBSITE PARA GESTORES

O Website será principalmente utilizado pelos administradores e gestores de escolas, para visualizar os gráficos e relatórios citados anteriormente. Exibindo estes dados e registros de forma

[1] biblioteca capaz de extrair o conteúdo textual de websites.

dinâmica, o sistema torna-se uma ferramenta para que os mesmos possam identificar as salas ou setores onde pesquisas de teor discriminatório estão ocorrendo com mais frequência, além da incidência e médias de bloqueio em cada área, entre outros.

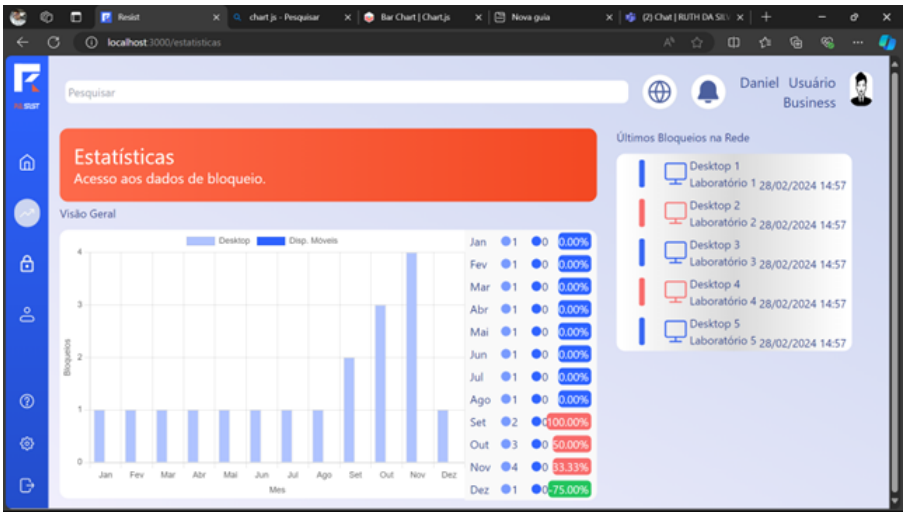
Figura 5 – Dashboard



Fonte: Autoria Própria (2024)

A Figura 5 ilustra a tela inicial, tratando-se de um *Dashboard* que proporciona um resumo das informações coletadas, como a atividade recente, que apresenta todas as atualizações do ambiente; visão geral dos bloqueios, trazendo a quantidade de bloqueios e a evolução em relação ao mês anterior; nível de incidência por laboratório (no exemplo, tratam-se de laboratórios de informática em um ambiente escolar); histórico por data, e os laboratórios que estão ativos no momento.

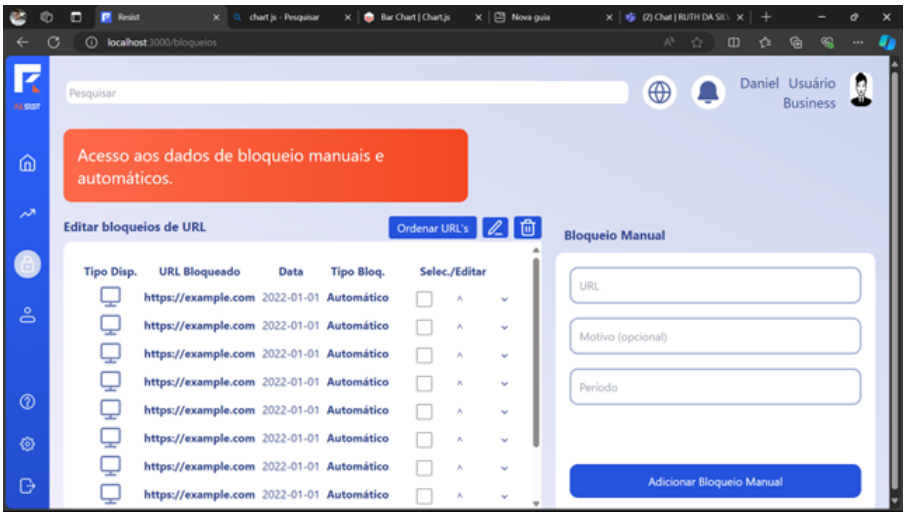
Figura 6 – Estatísticas



Fonte: Autoria Própria (2024)

A Figura 6 exibe a aba de estatísticas, onde há uma visualização aprofundada dos dados de bloqueio, exibindo a quantidade de bloqueios em dispositivos móveis e *Desktop*, uma comparação do aumento ou diminuição mês-a-mês, e alguns dos últimos bloqueios na rede.

Figura 7 – Bloqueios



Fonte: Autoria Própria (2024)

A página bloqueios, representada na Figura 7, exibe uma visão aprofundada sobre cada bloqueio realizado. É exibida uma tabela com a URL bloqueada, a data do bloqueio e se foi manual ou automático. A página disponibiliza ainda uma funcionalidade para bloquear um Website manualmente, especificando o motivo e período do bloqueio. É possível editar os bloqueios já realizados, para caso seja necessário cancelá-los.

Figura 8 – Notificação

Fonte: Autoria Própria (2024)

Assim que um site for bloqueado, uma notificação, exemplificada na Figura 8, aparece imediatamente na tela para o usuário.

INTELIGÊNCIA ARTIFICIAL

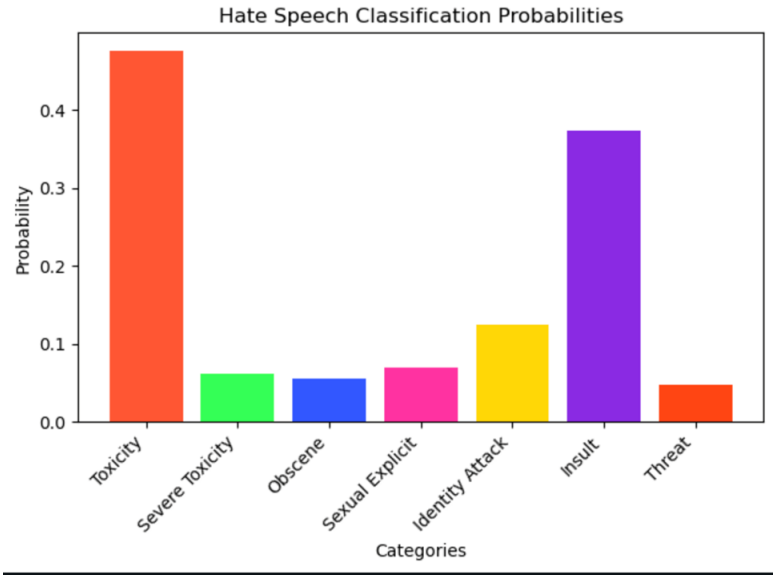
Atualmente, a IA ainda não está integrada ao sistema de monitoramento e bloqueio de sites e não foi treinada com 100% do *Dataset* devido a limitações no hardware utilizado. Para atingir resultados mais precisos e confiáveis, será necessário realizar treinamentos adicionais com um número

maior de amostras do dataset. No entanto, mesmo nessa fase inicial, a IA já é capaz de gerar gráficos que apresentam a probabilidade de determinada frase pertencer a uma das categorias específicas de discurso de ódio selecionadas, como demonstrado na Figura 9. Essa funcionalidade oferece uma visualização clara e detalhada do desempenho do modelo, destacando seu potencial para análises aprofundadas em contextos reais.

Embora a integração com o sistema de monitoramento esteja em fase de planejamento, a IA já apresenta resultados promissores. Quando integrada, substituirá o método atual de análise baseado em palavras-chave, oferecendo uma abordagem muito mais precisa e sofisticada, capaz de identificar discursos ofensivos mesmo em textos com linguagem implícita ou nuances. Essa integração também possibilitará a geração de relatórios detalhados, associando cada site bloqueado à categoria de discurso de ódio identificada, oferecendo maior controle e transparência para os gestores das instituições.

Com o uso dessa tecnologia, espera-se não apenas melhorar o monitoramento e bloqueio de conteúdos discriminatórios, mas também promover um ambiente digital mais seguro e respeitoso, especialmente em contextos sensíveis como escolas e outras instituições educacionais.

Figura 9 – Gráfico de classificação



Fonte: Autoria Própria (2024)

CONCLUSÃO

Este artigo destacou a aplicação da tecnologia e da Inteligência Artificial no contexto escolar, com foco na detecção e no bloqueio de acesso a sites com conteúdos discriminatórios. Diante dos crescentes desafios relacionados à inclusão étnica, educacional e social, a adoção de abordagens inovadoras se torna imprescindível para atender às demandas em constante evolução.

O projeto Resist está alinhado aos Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especificamente ao objetivo quatro, que visa garantir educação inclusiva, equitativa e de qualidade, e ao objetivo dez, que promove a inclusão social, econômica e política, independentemente de características como idade, sexo, raça, etnia, origem, religião ou condição econômica. A implementação do sistema transcende a simples detecção de conteúdos discriminatórios, abrangendo também o bloqueio de sites que contenham contextos de intolerância. Além disso, possibilita que instituições de ensino identifiquem tentativas de acesso a tais conteúdos, incentivando a conscientização sobre os impactos

desses discursos e promovendo uma cultura de respeito e diversidade. Desenvolvido em Python, o sistema extrai o conteúdo de sites, monitora os acessos registrados pelo Squid e retorna informações como URL, data, hora e IP da máquina.

Atualmente, a aplicação utiliza uma abordagem baseada na presença de palavras específicas para bloquear ou permitir o acesso, sendo limitada pela ausência de integração com a IA já desenvolvida. A URL da página é armazenada em um banco de dados para facilitar a verificação do status de liberação e identificar a palavra ou frase que motivou o bloqueio, se necessário. Após cada verificação, o sistema registra os acessos no banco de dados, permitindo a geração de relatórios detalhados, acessíveis em formato gráfico via Web pelos gestores das instituições. Embora ainda existam limitações, como a possibilidade de bloqueio indevido de conteúdos legítimos de cunho informativo, os resultados obtidos até o momento demonstram que o sistema representa um avanço significativo na utilização da tecnologia para promover valores fundamentais de igualdade, justiça e respeito mútuo.

O projeto adota um enfoque proativo e holístico, atacando as raízes do racismo e outros discursos de ódio, enquanto inspira esperança e estabelece um modelo para futuras iniciativas de inclusão. Para as melhorias futuras, destacam-se a integração do sistema com a IA treinada e o desenvolvimento de mecanismos de feedbacks contínuo para aprimorar a eficácia e a precisão das análises. Assim, o projeto não apenas se revela economicamente viável, mas também essencial para a construção de uma sociedade mais justa e equitativa, contribuindo diretamente para a promoção de uma educação inclusiva e de qualidade, alinhada aos ODS e aos valores de diversidade e inclusão.

REFERÊNCIAS

CANONICAL LTD. **Ubuntu Desktop**. [*S. l.: s. n.*]. Acesso em: 17 de abril 2024. Disponível em: <https://ubuntu.com/desktop>.

CASTAÑO-PULGARÍN, Santiago *et al.* Internet, social media and online hate speech. Systematic review. **Aggression and Violent Behavior**, Elsevier, v. 58, p. 101608, 2021.

DANTAS, MARCEL. **Que tal avaliar seu modelo? — Parte 1**. [*S. l.: s. n.*], 2019. Acesso em: 03 de fevereiro de 2023. Disponível em: <https://medium.com/data-hackers/que-tal-avaliar-seu-modelo-parte-1-66b24cbb8e7a>.

LEE, Hyun; KIM, Young-Han; KIM, Jung-Hye. **Report on Hate Speech**. [*S. l.*], 2019.

PELLE, Rômulo; MOREIRA, Viviane. Offensive Comments in Brazilian Web: a dataset and baseline results. *In*: PROCEEDINGS of the Brazilian Workshop on Social Network Analysis and Mining. [*S. l.: s. n.*], 2016. P. 510–519.

PELLE PELLE, Rogers Prates de; MOREIRA, Viviane P Moreira. Offensive Comments in the Brazilian Web: a dataset and baseline results. *In*: CONGRESSO da Sociedade Brasileira de Computação-CSBC. [*S. l.: s. n.*], 2017.

PITROPAKIS, Nikolaos *et al.* Monitoring users behavior: anti-immigration speech detection on twitter. **Machine Learning and Knowledge Extraction**, MDPI, v. 2, n. 3, p. 11, 2020.

POOJITHA, K *et al.* Classification of social media Toxic comments using Machine learning models, 2023.

SALEH, Hind; ALHOTHALI, Areej; MORIA, Kawthar. Detection of hate speech using bert and hate speech word embedding with deep model. **Applied Artificial Intelligence**, Taylor & Francis, v. 37, n. 1, p. 2166719, 2023.

SHARMA, Hitesh Kumar *et al.* Detecting hate speech and insults on social commentary using nlp and machine learning. **Int J Eng Technol Sci Res**, v. 4, n. 12, p. 279–285, 2017.

SRINIDHI, Sunny. Understanding word n-grams and n-gram probability in natural language processing. **Towards Data Science**, 2019.

SZYMANSKI, Heloisa. **A relação família-escola: desafios e perspectivas**. 2. ed. Brasília: Líber Livro, 2007.

THE SQUID SOFTWARE FOUNDATION. **Squid Cache**. [*S. l.: s. n.*]. Acesso em: 23 de abril 2024. Disponível em: <https://www.squid-cache.org/>.