PANDORA: Uma Plataforma Colaborativa para Transcrição Semiautomática de Boletins de Ocorrência Manuscritos

Wagner Santos[1,2], **Gabriel Coelho**[1], **Aline Paes**[1], **Isabel Rosseti**[1], **Daniel de Oliveira**[1]

[1]Universidade Federal Fluminense

[2]Secretaria Estadual de Polícia Militar do Rio de Janeiro
Rio de Janeiro - RJ - Brasil

{wagnergs, gabrielhcs}@id.uff.br, {alinepaes,rosseti,danielcmo}@ic.uff.br

Abstract. *The Police Report (PR) is one of the primary sources of information for the foundation and promotion of public security policies. Despite the existence of mobile applications for registering PRs, for multiple reasons, many police officers still register the PR in handwritten form. Registering the PR in handwriting is a challenge for collecting information, as it imposes a step of transcribing the text, which is an arduous and poorly scalable task. This paper proposes a collaborative platform, called* PANDORA*, which employs Machine Learning techniques to perform an initial transcription of the handwritten PRs to be modified/improved through the collaboration of multiple expert users. An evaluation with expert users and real PRs was performed.*

Resumo. *O Boletim de Ocorrência (BO) policial constitui uma das principais fontes de informação para a fundamentação e fomentação de políticas públicas de segurança. Apesar da existência de aplicativos móveis para registro dos BOs, por razões múltiplas muitos oficiais de polícia ainda registram o BO de forma manuscrita. O registo do BO de forma manuscrita é um desafio para a coleta de informações, pois impõe uma etapa de transcrição do texto, que é uma tarefa árdua e pouco escalável. Este artigo propõe uma plataforma colaborativa, denominada* PANDORA*, que utiliza técnicas de Aprendizado de Máquina para realizar uma transcrição inicial do BO manuscrito para então ser modificada/melhorada, por meio da colaboração de múltiplos usuários especialistas. Uma avaliação com usuários especialistas e BOs reais foi executada.*

1. Introdução

Nas últimas duas décadas, a população, em especial a dos grandes centros urbanos como o Rio de Janeiro, têm observado uma trajetória ascendente, mesmo que inconstante, na quantidade de ocorrências dos chamados "crimes comuns" como roubo/furto de pequenos bens (*e.g.*, telefone celular), roubo/furto de bens maiores (*e.g.*, automóveis), entre outros delitos [Kopittke and Ramos 2021]. Sempre que uma ocorrência de atividade criminal é relatada por um cidadão (ou por um grupo), um documento comprobatório deve ser elaborado para registrar formalmente a ocorrência: o *Boletim de Ocorrência* (BO). Tais boletins constituem uma das principais fontes de informação para a fundamentação e fomentação de políticas públicas com foco na redução da quantidade de ocorrências criminais [Lourenço et al. 2018, Alikhademi et al. 2022].

Por mais que existam variações na estrutura do BO em cada estado brasileiro, em geral, eles são tradicionalmente compostos por múltiplos campos que identificam o cidadão que foi alvo de uma ação criminosa, além de dados da própria ocorrência, incluindo endereço, tipo de crime, horário, descrição dos fatos, *etc*. Os BOs são a principal fonte para a identificação de *Hot Spots* criminais [1] [Reis et al. 2006] (*i.e.*, áreas da cidade onde o índice de criminalidade é mais elevado) e para elaboração de políticas de segurança pública. Entretanto, durante anos, os BOs foram somente preenchidos de forma manual pelos policiais militares, o que acabava dificultando que o seu conteúdo fosse considerado nas estatísticas oficiais, uma vez que o mesmo dependia de transcrição. Mesmo nos dias de hoje, em que boa parte das secretarias de segurança possui aplicativos móveis para registro do BO, quando não há cobertura de sinal de internet no local da ocorrência ou quando há ausência de equipamentos, o boletim ainda é manuscrito.

Na grande maioria dos casos, as transcrições dos BOs manuscritos são realizadas por especialistas da área de segurança pública que identificam o conteúdo do BO e o representam em formato digital. Pode-se perceber que essa é uma tarefa árdua, propensa a erros (dependendo da grafia do oficial de polícia) e pouco escalável, uma vez que o volume de BOs comumente supera em várias ordens de grandeza a quantidade de transcritores. Ainda, não é incomum que diversos transcritores tenham que trabalhar em um mesmo BO até que a transcrição final seja disponibilizada. Entretanto, com o desenvolvimento da área de Reconhecimento de Texto Manuscrito (HTR, do inglês *Handwritten Text Recognition*) [Plamondon and Srihari 2000], diversas técnicas e métodos têm sido propostos para permitir que modelos treinados em um determinado idioma e contexto gerem uma transcrição inicial que os especialistas podem alterar *a posteriori*, obtendo assim uma maior produtividade na tarefa de transcrição.

Ao mesmo tempo em que a área de HTR evoluiu, o uso de plataformas colaborativas para multidões (*i.e.*, *crowdsourcing*), onde diversos indivíduos são capazes de contribuir para uma determinada tarefa, se tornou popular para auxiliar na transcrição de textos manuscritos em diversos domínios. Entretanto, plataformas de *crowdsourcing* genéricas, como o *Amazon Mechanical Turk*[2], ou específicas, como o AnnoTate[3], não podem ser usadas para apoiar a transcrição de BOs. Em primeiro lugar porque tais plataformas são de uso público e hospedadas fora do território brasileiro, e os BOs são documentos que possuem informação sensível. Além disso, os BOs comumente apresentam caligrafia pouco elaborada e termos específicos do jargão policial, o que dificulta a obtenção de transcrições de qualidade realizadas por não especialistas. Finalmente, por se tratar de um domínio altamente especializado, o modelo treinado de HTR que processa o texto manuscrito deve ser refinado para o domínio, uma vez que não é trivial desenvolver um sistema genérico que possa processar todos os tipos de textos, linguagens e grafias [Purohit and Chauhan 2016].

Sendo assim, considerando as características específicas da transcrição de BOs manuscritos e a utilização de sistemas de informação e colaboração para transcrição de documentos manuscritos já existentes [Granell and Martínez-Hinarejos 2016], este artigo apresenta a PANDORA (**P**lataform**A** para transcrição de boleti**N**s **D**e **O**cor**R**ência m**A**nuscritos), uma plataforma colaborativa para apoiar a tarefa de transcrição semiautomática de BOs

[1] Termo utilizado para se referir a áreas geográficas que apresentam uma concentração anormal de atividades criminosas, tais como furtos, roubos, tráfico de drogas, entre outros.

[2] https://www.mturk.com/

[3] https://anno.tate.org.uk/

manuscritos. A PANDORA utiliza técnicas de HTR e tem como objetivo captar BOs manuscritos, convertê-los para formato digital, realizar ajustes na conversão de forma colaborativa e utilizar mecanismos de eleição para definir a melhor transcrição possível. A plataforma foi avaliada por meio de um estudo piloto que envolveu a importação de 2.000 BOs manuscritos oriundos de diversas Unidades Policiais Militares distribuídas pelo estado do Rio de Janeiro. Desses, 156 foram avaliados de forma colaborativa.

O presente artigo está organizado em quatro seções além desta introdução. A Seção 2 apresenta conceitos importantes e os trabalhos relacionados. A Seção 3 apresenta detalhes da plataforma PANDORA. A Seção 4 apresenta a avaliação da proposta e, na Seção 5, são apresentadas as conclusões e listados trabalhos futuros.

2. Referencial Teórico e Trabalhos Relacionados

Nesta seção, são apresentados conceitos necessários para compreensão da abordagem proposta, os quais envolvem Reconhecimento de Texto Manuscrito (HTR) (Seção 2.1) e a estrutura do Boletim de Ocorrência (Seção 2.2). Em seguida, na Seção 2.3, são elencados trabalhos na literatura que tem relação com a plataforma proposta neste artigo.

2.1. Aprendizado de Máquina e HTR

A área de Aprendizado de Máquina [Mitchell and Mitchell 1997] é uma subárea da Inteligência Artificial que tem como objetivo propor algoritmos que aprendam a resolver tarefas sem que tenham sido explicitamente programados. Exemplos bem conhecidos de aplicação de aprendizado de máquina são a previsão de perfis de clientes em lojas e a classificação de clientes que solicitam empréstimos bancários [Mitchell and Mitchell 1997]. Essas aplicações tradicionalmente consomem dados estruturados para realizar o aprendizado. Porém, esse não é o cenário de aplicações que consomem textos manuscritos, em que é inviável estimar todos as formas possíveis que uma determinada letra pode ser escrita. Assim, o uso de modelos de Aprendizado de Máquina se apresenta como uma solução capaz de possibilitar a identificação automática dos caracteres e, consequentemente, das palavras que compõem um texto manuscrito.

O aprendizado baseado em textos manuscritos (HTR) comumente se vale do aprendizado de máquina supervisionado [Mitchell and Mitchell 1997]. Nesse caso, o método recebe como entrada um conjunto de imagens com textos manuscritos, pareadas com os textos em formato digital, o que constitui o conjunto de exemplos. Durante o aprendizado, o método induz um modelo ajustando-o de forma a minimizar o erro entre o texto manuscrito identificado na imagem pelo modelo e o respectivo texto digital fornecido como entrada [Bonaccorso 2018]. Portanto, um sistema baseado em HTR é composto por um modelo preditivo treinado a partir de um aprendizado supervisionado e que é capaz de inferir as informações textuais contidas nas imagens de documentos manuscritos fornecidas como entrada do sistema.

Para que as imagens de entrada possam servir de exemplos para o aprendizado, é necessário que elas passem por um processo de pré-processamento. As etapas para efetuar o pré-processamento, neste caso, envolvem: (i) a aplicação de filtros, para aumentar o contraste entre o texto e o restante da imagem, (ii) a binarização do valor dos *pixels* para 0 ou 1, alterando a imagem para tons de branco e preto, respectivamente, (iii) a detecção de linhas para identificar componentes conectados entre os elementos nos pixels, e (iv) o

reconhecimento textual, em que se associa as informações obtidas dos contornos presentes nos *pixels* a símbolos provenientes do meio textual usado como base. A execução dessas etapas é crucial para que o método de aprendizado receba uma entrada reduzida de ruídos.

O aprendizado de um modelo de HTR, em geral, é composto por duas etapas principais: (i) a detecção de regiões do texto, seguida pela (ii) transcrição das informações textuais contidas nas respectivas regiões identificadas. Em geral, o primeiro passo se vale de métodos de processamento de imagens, mas também tem sido abordado sob o ponto de vista de Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Networks*) [Khan et al. 2018], em que os modelos são treinados para extrair características nas imagens que identifiquem as informações contidas nas mesmas como texto. O segundo passo, em geral, é executado utilizando Redes Neurais Recorrentes (RNN do inglês *Recurrent Neural Networks*) [Bezerra et al. 2012], responsáveis por processar informações que apresentam caráter sequencial [Bezerra et al. 2012].

2.2. Boletim de Ocorrência

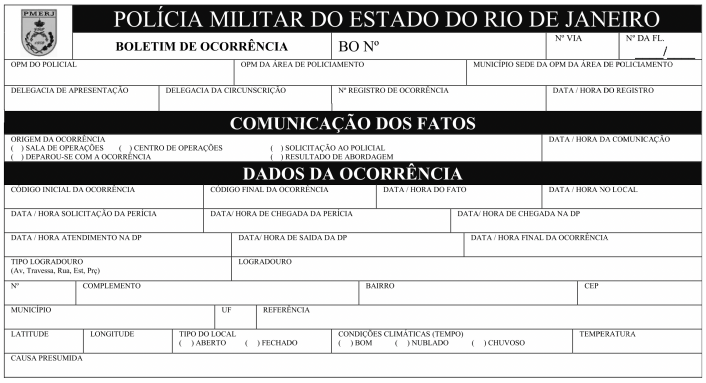
A emissão de BOs por parte de policiais militares em todos os estados do Brasil tem previsão legal no §3 do artigo 5º do Código de Processo Penal, cuja redação menciona que *"Qualquer pessoa do povo que tiver conhecimento da existência de infração penal em que caiba ação pública poderá, verbalmente ou por escrito, comunicá-la à autoridade policial, e esta, verificada a procedência das informações, mandará instaurar inquérito"*, servindo para registrar ocorrências criminais. Especialmente no caso do Rio de Janeiro, o BO é o principal documento operacional produzido pela Polícia Militar. Em virtude de sua importância, diversas instituições públicas e privadas solicitam seus dados com amparo da Lei nº 12.527, sancionada em 18 de novembro de 2011 (Lei de Acesso à Informação).

A Polícia Militar do Estado do Rio de Janeiro (PMERJ) define que o BO é um instrumento formal destinado ao registro de dados de ocorrências criminais e assistenciais e dos registros de reuniões de mediação de conflitos e conselhos comunitários que sejam presididos/mediados por policiais militares [SEPM 2015]. Embora o código de processo penal faça menção à comunicação da existência de infração penal em que caiba a ação pública, o BO tem como principal objetivo oferecer à sociedade civil a transparência relativa necessária nas ações realizadas pela Polícia Militar quando em contato com a população.

Cada BO tem uma estrutura bem definida, que serve para atender às demandas dos órgãos de controle, bem como da Secretaria Nacional de Segurança Pública – (SENASP). O BO definido pela PMERJ é composto por 156 campos divididos em seções bem definidas que oferecem a possibilidade do acompanhamento cronológico de todas as etapas realizadas pelo policial militar durante o atendimento ao público, a identificação de todos os envolvidos (autor, testemunha, vítimas e solicitantes) e da equipe de serviço, além de toda a narrativa, contextualizada e de forma resumida, dos fatos ocorridos ou de acordos firmados nas reuniões de mediação de conflitos e nos conselhos comunitários. A Figura 1 apresenta um fragmento da estrutura de um BO da Polícia Militar do Estado do Rio de Janeiro.

2.3. Trabalhos Relacionados

Alguns trabalhos encontrados na literatura discutidos a seguir apresentam plataformas para transcrição de texto manuscrito. Podemos identificar duas categorias: (i) plataformas independentes de domínio e (ii) plataformas específicas de um domínio de aplicação.



POLÍCIA MILITAR DO ESTADO DO RIO DE JANEIRO

BOLETIM DE OCORRÊNCIA | BO Nº | Nº VIA | Nº DA FL.

OPM DO POLICIAL | OPM DA ÁREA DE POLICIAMENTO | MUNICÍPIO SEDE DA OPM DA ÁREA DE POLICIAMENTO

DELEGACIA DE APRESENTAÇÃO | DELEGACIA DA CIRCUNSCRIÇÃO | Nº REGISTRO DE OCORRÊNCIA | DATA / HORA DO REGISTRO

COMUNICAÇÃO DOS FATOS

ORIGEM DA OCORRÊNCIA
( ) SALA DE OPERAÇÕES ( ) CENTRO DE OPERAÇÕES ( ) SOLICITAÇÃO AO POLICIAL
( ) DEPAROU-SE COM A OCORRÊNCIA ( ) RESULTADO DE ABORDAGEM
DATA / HORA DA COMUNICAÇÃO

DADOS DA OCORRÊNCIA

CÓDIGO INICIAL DA OCORRÊNCIA | CÓDIGO FINAL DA OCORRÊNCIA | DATA / HORA DO FATO | DATA / HORA NO LOCAL

DATA / HORA SOLICITAÇÃO DA PERICIA | DATA/ HORA DE CHEGADA DA PERICIA | DATA/ HORA DE CHEGADA NA DP

DATA / HORA ATENDIMENTO NA DP | DATA/ HORA DE SAIDA DA DP | DATA / HORA FINAL DA OCORRÊNCIA

TIPO LOGRADOURO (Av, Travessa, Rua, Est, Prç) | LOGRADOURO

Nº | COMPLEMENTO | BAIRRO | CEP

MUNICÍPIO | UF | REFERÊNCIA

LATITUDE | LONGITUDE | TIPO DO LOCAL ( ) ABERTO ( ) FECHADO | CONDIÇÕES CLIMÁTICAS (TEMPO) ( ) BOM ( ) NUBLADO ( ) CHUVOSO | TEMPERATURA

CAUSA PRESUMIDA

Figura 1. Fragmento de um Boletim de Ocorrência da PMERJ

Silva (2021) apresenta um *pipeline* para transcrição automática de textos manuscritos em português baseado em técnicas de Aprendizado de Máquina. O *pipeline* explora métodos e ferramentas consolidadas para extração e reconhecimento de informação textual em imagens. A análise realizada por Silva (2021) apontou incongruências nos aspectos relacionados ao estilo de escrita empregado e às particularidades da língua portuguesa, como o uso de símbolos, acentuação e combinação de palavras.

Linkforman-Sulen *et al.* (1995) propõem um método capaz de detectar linhas de texto manuscritas em várias direções em uma mesma página, inclusive rasuras ou anotações entre linhas. O método possui uma estratégia de validação de hipóteses que é ativada iterativamente até que o final da segmentação seja alcançado. A cada estágio, o método analisa a melhor hipótese para geração da linha de texto, levando em consideração as flutuações dos componentes da linha de texto. Em seguida, o modelo verifica a validade da linha no domínio da imagem por meio de um critério de proximidade, analisando o contexto em que se percebe o alinhamento.

Louloudis *et al.* (2009) apresentam um método de detecção de linhas de texto para documentos manuscritos. O método proposto é baseado em uma estratégia que consiste em quatro etapas distintas: (i) a binarização e aprimoramento da imagem, (ii) a extração de componentes conectados, (iii) o particionamento do domínio do componente conectado em três subdomínios espaciais e (iv) a estimativa da altura média dos caracteres. A transformada de Hough é usada para a detecção de possíveis linhas de texto, para corrigir possíveis divisões na detecção de linhas de texto, e, finalmente, na separação de linhas conectadas. É importante mencionar que as abordagens supracitadas não focam em colaboração, somente na transcrição inicial realizada por métodos de Aprendizado de Máquina.

Granell *et al.* (2018) propõem uma plataforma para a transcrição de textos históricos. A abordagem proposta realiza uma fusão de dados considerando textos manuscritos e áudio de descrição desses textos. A plataforma proposta por Granell *et al.* (2018) trabalha de forma colaborativa e permite que diversos usuários especialistas em transcrições de documentos históricos possam ajustar a transcrição inicial identificada pelo modelo de HTR. Similarmente, Tomic *et al.* (2021) propõem uma plataforma para transcrição de documentos históricos croatas que seguem o alfabeto glagolítico. Nessa plataforma, não há utilização de métodos de HTR, somente a exploração da força de trabalho de cidadãos que conhecem o alfabeto e se dispõem a traduzir tais documentos.

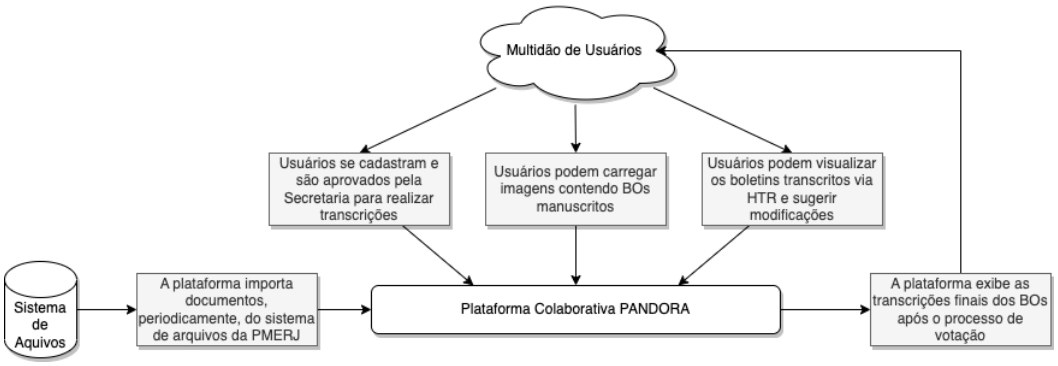


Figura 2. Diagrama de Contexto do processo de colaboração da PANDORA.

3. A Plataforma Colaborativa PANDORA

A plataforma PANDORA tem como objetivo prover um ambiente colaborativo para a transcrição semiautomática de BOs manuscritos. A Figura 2 apresenta um diagrama de contexto que ilustra as responsabilidades, processos e interações da PANDORA com os elementos externos relacionados, *i.e.*, usuários que se encontram aptos a trabalhar no processo de transcrição de BOs. Este diagrama facilita a compreensão do sistema e apoia o levantamento de requisitos.

A PANDORA foi projetada como uma aplicação *Web* de modo que possa alcançar um grande número de usuários em virtude do acesso restrito de dispositivos móveis à intranet da PMERJ. Como os dados tratados são sensíveis, a plataforma só permite a colaboração de usuários previamente cadastrados e aprovados pelo órgão de segurança competente, *i.e.*, PMERJ. Conforme mencionado anteriormente, as Secretarias de Segurança já disponibilizam aplicativos para dispositivos móveis para registro de BOs. Porém, problemas operacionais (*e.g.*, quando há falta de cobertura de internet no local da ocorrência) acabam induzindo que os policiais preencham manualmente o BO e o encaminhe ao Centro de Atendimento de Emergência – CAE-190. Esses formulários são acumulados e enviados, ao término do serviço, à seção responsável para que outros policiais e funcionários civis possam digitalizar os documentos manuscritos. Essas imagens são armazenadas em uma área específica do sistema de arquivos da PMERJ.

A partir das imagens dos BOs manuscritos armazenadas no sistema de arquivos, a PANDORA importa BOs em bloco para serem transcritos via HTR e revisados pelos usuários. Os usuários autorizados podem também carregar os BOs para processamento diretamente no portal *Web* da plataforma, sem utilizar a carga automática via sistema de arquivos. Uma vez carregados, os BOs entram em uma fila de processamento e são geradas as transcrições iniciais via HTR. Com os BOs uma vez processados, os usuários podem então sugerir modificações nas transcrições iniciais de acordo com sua experiência no domínio da aplicação e conhecimento do vocabulário específico. No caso em que múltiplas sugestões são realizadas, a PANDORA possui um mecanismo de votação para decidir a transcrição final. Os documentos transcritos ficam disponíveis para sugestões durante uma janela de tempo. Uma vez que uma votação é realizada e a transcrição final é gerada, não é mais possível sugerir modificações para a transcrição. A colaboração por meio de *crowdsourcing* enriquece a plataforma devido ao caráter heterogêneo dos seus usuários, *e.g.*, policiais militares, funcionários civis especialistas em segurança, sociólogos, *etc*. Entretanto, o uso de *crowdsourcing* deve ser guiado por mecanismos de governança da plataforma [Blohm et al. 2018],

conforme discutido a seguir.

3.1. Mecanismos de Governança para *Crowdsourcing* na PANDORA

Diversos mecanismos de governança podem ser elencados de acordo com o tipo de plataforma. Apesar de existirem diferentes classificações de plataformas de *crowdsourcing*, neste artigo seguimos a classificação definida por [Faber et al. 2018] em que a plataforma PANDORA pode ser classificada como uma plataforma de *crowdsourcing* para solução de problemas complexos. Assim, inspirados na definição de governança de [Alves et al. 2022], a Tabela 1 apresenta alguns mecanismos de governança e como foram tratados no contexto da plataforma PANDORA.

Tabela 1. Mecanismos de Governança da plataforma PANDORA

Mecanismo	Característica	Tratamento
Requisito de Contribuição	Dados fundamentais para que a colaboração possa acontecer na plataforma PANDORA.	Obtenção de imagens contendo os BOs manuscritos.
Alocação de Tarefas	Como as tarefas de transcrição são distribuídas pelos usuários, de forma a evitar alocações tendenciosas que diminuam a qualidade da transcrição final.	Limitação de transcrições por usuário por tipo de ocorrência (*e.g.*, homicídio) e por área de cobertura de um batalhão.
Recompensa	Mecanismo de recompensa utilizado para motivar a participação dos usuários.	Ranking divulgado na plataforma com a quantidade de contribuições por usuário que foram escolhidas para a transcrição final.
Tutorial/Treinamento	Disponibilização de documentos com passo-a-passo para uso.	Elaboração de manuais e videos que explicam o uso da PANDORA, além de treinamento *in loco*.

3.2. Arquitetura da Plataforma PANDORA

Conforme mencionado anteriormente, a PANDORA é um plataforma que visa permitir a transcrição colaborativa de BOs, e sua arquitetura é apresentada na Figura 3. A arquitetura proposta é organizada em seis componentes principais: (i) *Broker* de Dados, (ii) Portal *Web*, (iii) Gerenciador de Fila, (iv) Processador HTR, (v) *Data Lake* e (vi) Gerenciador de Votação. A seguir apresentamos detalhes de cada um dos componentes.

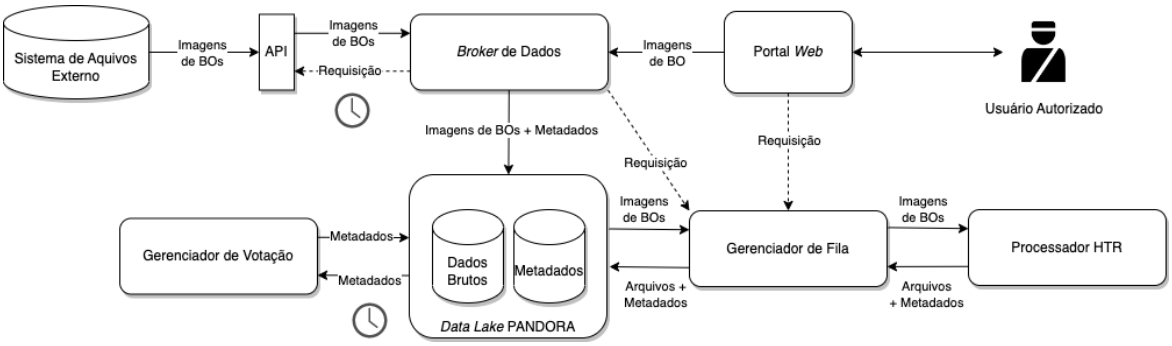


Figura 3. Arquitetura da Plataforma PANDORA

A plataforma obtém as imagens dos BOs manuscritos de duas formas: (i) via API que acessa de forma agendada o sistema de arquivos da PMERJ onde se encontram os BOs

manuscritos e (ii) via carga direta pelo Portal *Web*. Em ambos os casos o *Broker* de Dados é invocado. Este componente é responsável por acessar os arquivos brutos com as imagens e armazená-los no *Data Lake* da PANDORA em diretórios específicos. Na PANDORA, seguimos a organização de diretórios de acordo com o valor *hash* SHA-1 do conteúdo de cada arquivo. Um primeiro nível de diretórios no *Data Lake* é criado com os quatro primeiros caracteres do SHA-1. Um segundo nível com os quatro caracteres seguintes, e, finalmente, o nome do arquivo são os 32 caracteres restantes. Esta organização evita o armazenamento duplicado de arquivos idênticos, e também a criação de uma quantidade exagerada de diretórios. Além dos arquivos brutos, o *Broker* de Dados também armazena uma série de metadados associados a cada arquivo, *e.g.*, tipo de arquivo, qual número do BO associado, *etc*. A Figura 4 apresenta o modelo de dados do repositório de metadados.

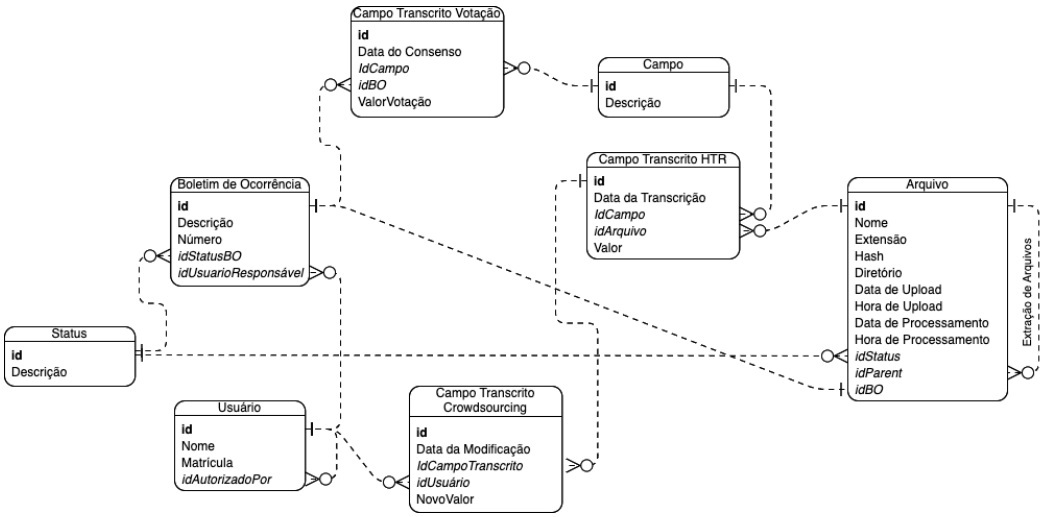


Figura 4. Modelo de Metadados da Plataforma PANDORA

No modelo apresentado na Figura 4 é possível perceber que cada *Boletim de Ocorrência* se encontra associado a um *Arquivo*. Cada *Arquivo* possui múltiplos *Campos* que fazem parte de um BO. Após o processo de HTR, um arquivo contendo a imagem de um BO é descomposto em múltiplos arquivos (vide explicação componente Processador HTR), que devem estar associados. Uma vez que o processo de HTR seja executado, os valores de cada campo são armazenados na tabela *Campo Transcrito HTR*. Cada *Usuário* da PANDORA pode sugerir modificações nos valores dos campos que foram transcritos inicialmente pelo processo de HTR. Cada sugestão de cada campo é armazenada na Tabela *Campo Transcrito Crowdsourcing*. Finalmente, ao final da etapa de votação, a Tabela *Campo Transcrito Votação* é preenchida com os valores dos campos da transcrição definitiva.

Uma vez que os arquivos com as imagens dos BOs manuscritos estejam no *Data Lake* e seus metadados armazenados no repositório, o componente Gerenciador de Fila é invocado. Este componente é o responsável por coordenar o processamento HTR de cada arquivo carregado na PANDORA. O Gerenciador de Fila invoca o Processador HTR para cada um dos arquivos que se encontram na fila de processamento. No Processador HTR, seguimos o *pipeline* definido por Silva (2021), onde é utilizada a ferramenta CRAFT de detecção de textos em imagens [Baek et al. 2019]. Cada imagem contendo informação textual dada como entrada é processada por meio de uma CNN, e são gerados quatro arquivos de saída, *i.e.*, a imagem com as regiões que tem palavras demarcadas, um arquivo ASCII

contendo as coordenadas das regiões identificadas, um mapa de calor que indica os pontos mais relevantes onde estão contidos os caracteres, e o outro mapa de calor que indica os espaços entre os pontos mencionados anteriormente.

A seguir, é invocada a biblioteca OpenCV que recorta as áreas correspondentes das palavras segmentadas, gerando novas imagens contendo uma palavra demarcada em cada. É importante ressaltar que cada arquivo gerado a partir do processo de HTR para uma imagem de entrada é gravado no *Data Lake* e seus metadados no repositório de metadados. Cada imagem extraída é então dada como entrada para a ferramenta de reconhecimento de textos SimpleHTR [Nikitha et al. 2020] que, com uma arquitetura híbrida composta por camadas CNN e RNN, identifica a palavra presente na imagem, retornando-a juntamente com a probabilidade de corresponder à palavra correta.

Finalmente, depois que os usuários enviam suas sugestões de modificação para os campos transcritos pelo processo de HTR, o Gerenciador de Votação é invocado após uma janela de tempo configurável. O Gerenciador de Votação na versão atual segue o critério da maioria, *i.e.*, vence a sugestão que tiver obtido a maioria simples das modificações. É importante ressaltar que mecanismos mais elaborados de votação podem ser aplicados ou mecanismos que levam em consideração o nível de experiência de cada usuário. A versão beta da PANDORA se encontra disponível para *download* em `https://github.com/UFFeScience/pandora`.

4. Avaliação da Plataforma PANDORA

Esta seção apresenta uma avaliação da plataforma PANDORA realizada por um conjunto de usuários especialistas, de modo a verificar a viabilidade da solução proposta. O cenário de uso escolhido representa o trabalho diário da equipe de transcrição e foca em avaliar como a colaboração pode ser realizada por meio desta nova plataforma.

Para o processo de avaliação da PANDORA foram solicitados à PMERJ 2.000 BOs oriundos de diversas Unidades Policiais Militares espalhadas pelo Estado em formato PDF. A partir de uma análise manual do conjunto de BOs, notou-se que muitos possuíam algum nível de imperfeição (Figura 5), tais como: rasuras, manchas brancas sobre as letras, anotações incompatíveis com o conteúdo do BO registradas nas bordas do documento ou sobre a informação nos campos (sinais de carimbos, *etc*.), distorções na imagem, dentre outras. Tais imperfeições levam a problemas na transcrição automática realizada pelo processo de HTR, e consequentemente, aumentam a quantidade de sugestões/modificações que deveriam ser realizadas pelos usuários. Do total de 2.000 BOs, somente 156 não apresentaram imperfeições, sendo processados em sua completude pelo processo de HTR.

O uso da plataforma PANDORA é iniciado quando o usuário carrega um arquivo contendo as páginas do BO ou com a importação automática via API. A plataforma registra o arquivo na fila para processamento. Para a identificação do arquivo na PANDORA, o usuário deve informar o número do BO associado ao arquivo que está sendo carregado. Uma vez carregado, o BO aparece na fila de processamento junto com o seu *status* atual de processamento ("*Não processado*", "*Em Processamento*", "*Processado*" e "*Processado com Erro*"), conforme exemplificado na Figura 6a. Uma vez que o BO tenha sido processado, o usuário pode selecioná-lo e visualizar o conteúdo transcrito pelo processo de HTR (Figura 6b) e sugerir modificações para cada um dos campos do BO.

De forma a avaliar a plataforma PANDORA, utilizamos o modelo de avaliação deno-

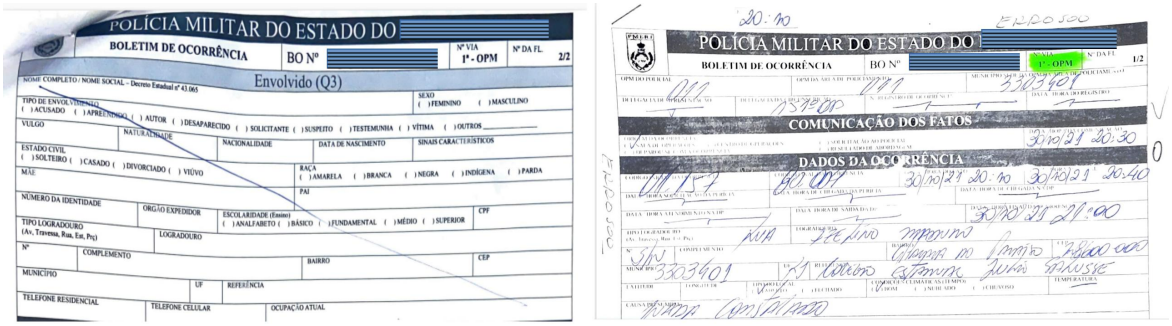


Figura 5. Exemplos de BOs com problemas de digitalização.

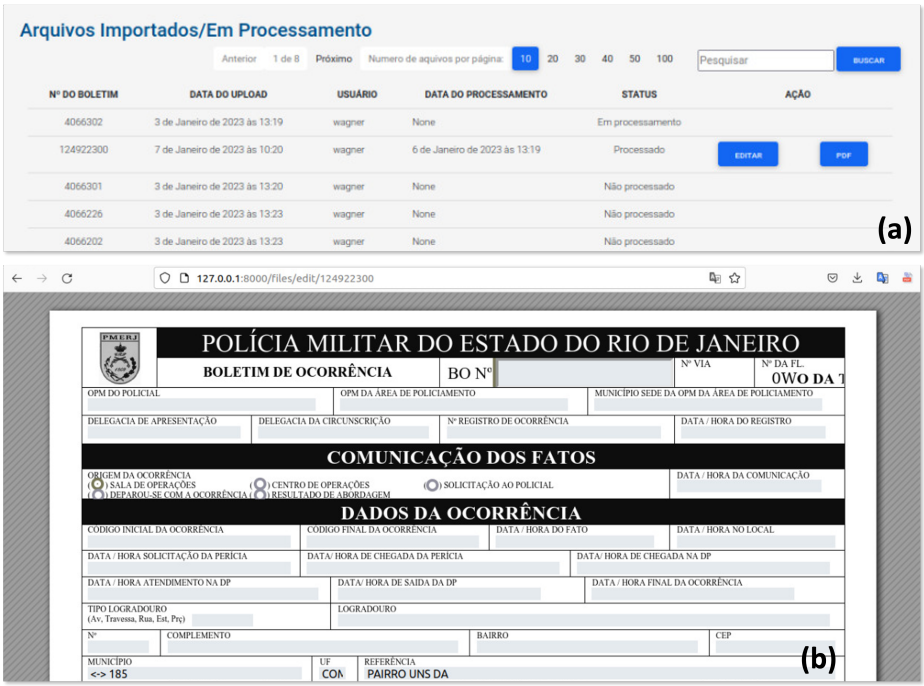


Figura 6. Interface da Plataforma PANDORA para (a) Visualização da Fila de Processamento de BOs e (b) Sugestão de Modificações da Transcrição Inicial.

minado TAM *(Technology Acceptance Model)* [Davis 1989]. A ideia principal do TAM é avaliar a receptividade/comportamento de um usuário no que se refere à facilidade e utilidade da tecnologia/ferramenta que está sendo proposta, neste caso a PANDORA. A utilidade refere-se ao quanto o usuário acredita que a plataforma proposta o auxiliará em suas tarefas, e a facilidade se refere ao quão simples será utilizar tal plataforma. A avaliação contou com seis usuários especialistas nas áreas de segurança pública e de análise de dados criminais e ocorreu no período compreendido entre Dezembro/2022 e Janeiro/2023 presencialmente na Secretaria de Polícia Militar do Rio de Janeiro. Os usuários escolhidos são todos policiais experientes ou funcionários civis especialistas em segurança pública, que conhecem o domínio da aplicação e os jargões utilizados nos documentos manuscritos. É importante ressaltar que, por se tratar de dados sensíveis e domínio especializado, não é possível utilizar usuários não-especialistas na avaliação. As formações de cada especialista são apresentadas na Tabela 2.

As avaliações para as questões do TAM são apresentadas na Tabela 3. É possível

Tabela 2. Formação dos avaliadores

ID Participante	Formação	Experiência na Função
1	Graduação em Sistema de Informação; Formação em metodologias quantitativas	08 anos
2	Graduação em Direito; Formação em Metodologia quantitativa	06 anos
3	Tecnólogo em Segurança Pública e Formação em metodologias quantitativas	12 anos
4	Graduação em Tecnologia em Segurança Pública	14 anos
5	Tecnólogo em Segurança Pública; Pós-graduação em direito penal; Formação em metodologias quantitativas	14 anos
6	Graduação em Tecnologia em Segurança Pública; Pós-graduação em direito penal	14 anos

perceber que os usuários tiveram uma boa percepção da plataforma PANDORA, apesar de 20% dos usuários terem apontado que a qualidade dos resultados pode ser baixa por conta de falhas no processo de HTR. Esta crítica já era esperada, uma vez que o modelo de HTR não foi treinado com os BOs, mas apenas com textos genéricos. Algumas correções de erro e melhorias também foram apontadas. O primeiro ponto foi a possibilidade de uma exclusão lógica de BOs, sem que o arquivo fosse excluído do *Data Lake*. Outro ponto de melhoria é a criação de perfis hierárquicos de usuários, cada um com um nível específico de visibilidade dos BOs. Apesar das sugestões de melhorias, a versão atual da PANDORA se mostrou promissora no que tange à transcrição de documentos manuscritos, em especial em um cenário tão especializado quanto o de segurança pública. A plataforma PANDORA se encontra em processo de homologação/teste para ser incorporada ao *pool* de sistemas da PMERJ em produção.

Tabela 3. Resultado da Avaliação da plataforma PANDORA com o TAM

Pergunta	Muito Baixo	Baixo	Médio	Alto	Muito Alto
Escolha o grau da aplicabilidade da PANDORA no seu trabalho.	0%	0%	0%	20%	80%
Escolha o grau do desempenho que seria obtido com a aplicação da PANDORA no seu trabalho.	0%	0%	20%	20%	60%
Escolha o grau da qualidade das informações exibidas pela PANDORA.	0%	60%	40%	0%	0%
Escolha o grau da qualidade dos resultados que você poderia obter ao usar PANDORA.	0%	20%	0%	60%	20%
Escolha o grau da facilidade de uso da PANDORA.	0%	0%	40%	20%	40%
Escolha o grau da facilidade de aprendizado com pouco ou nenhum treinamento da PANDORA.	0%	0%	0%	80%	20%
Escolha o grau da facilidade de lembrar de como se usa a PANDORA.	0%	0%	0%	80%	20%
Escolha o grau da facilidade de identificar quando ocorrem erros no PANDORA.	0%	20%	20%	40%	20%

5. Conclusões e Trabalhos Futuros

Este artigo apresentou uma plataforma colaborativa para apoiar o processo de transcrição de BOs manuscritos chamada PANDORA. Para tanto, a plataforma incorpora métodos para identificação automática de textos manuscritos baseado em Aprendizado de Máquina. O texto resultante do processo de HTR pode ser modificado colaborativamente pelos usuários especialistas. A PANDORA contribui para o desenvolvimento de sistemas colaborativos baseados em *crowdsourcing* no contexto de segurança pública.

Para demonstrar a viabilidade da plataforma, apresentamos uma avaliação com a importação de 2.000 BOs reais e a avaliação de 156 deles, explicando as funcionalidades atuais da plataforma e o contexto de uso planejado. Apesar de ter sido bem recebida pelos usuários, a partir dos resultados obtidos na avaliação com usuários especialistas, percebemos que, refinando o treinamento realizado com o modelo de HTR (*i.e.*, redes neurais), a plataforma PANDORA seria capaz de auxiliar de forma mais significativa na tarefa de transcrição, uma vez que com uma transcrição automática de maior qualidade, os usuários necessitariam realizar menos modificações *a posteriori*. Assim, um próximo passo para o projeto é o refinamento do modelo de HTR para o domínio de segurança pública, de forma que a transcrição inicial produza resultados de melhor qualidade.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Os autores também gostariam de agradecer ao CNPq (311275/2020-6, 311898/2021-1 e 316625/2021-3) e a FAPERJ (E-26/202.806/2019, E-26/202.914/2019, SEI-260003/000614/2023) pelo apoio financeiro.

Referências

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., and Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artif. Intell. Law*, 30:1–17.

Alves, R., David, J., Braga, R., Siqueira, K., Stroele, V., Barbosa, G., da Costa, J. P., and da Silva, I. (2022). Nutrinprice: uma plataforma colaborativa para apoiar a seleção de produtos alimentícios. In *SBSC*, pages 9–16. SBC.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character region awareness for text detection. In *IEEE CVPR*, pages 9365–9374. IEEE.

Bezerra, B. L. D., Zanchettin, C., and Braga de Andrade, V. (2012). A hybrid rnn model for cursive offline handwriting recognition. In *SBRN*, pages 113–118.

Blohm, I., Zogaj, S., Bretschneider, U., and Leimeister, J. M. (2018). How to manage crowdsourcing platforms effectively? *Calif. Manag. Rev.*, 60(2):122–149.

Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.*, 13(3):319–340.

Faber, A., Rehm, S., Hernandez-Mendez, A., and Matthes, F. (2018). Collectively constructing the business ecosystem: Towards crowd-based modeling for platforms and infrastructures. In *ICEIS*, pages 158–172.

Granell, E. and Martínez-Hinarejos, C. D. (2016). A multimodal crowdsourcing framework for transcribing historical handwritten documents. In *DocEng*, pages 157–163. ACM.

Granell, E., Romero, V., and Martínez-Hinarejos, C. D. (2018). Multimodality, interactivity, and crowdsourcing for document transcription. *Comput. Intell.*, 34(2):398–419.

Khan, S. H., Rahmani, H., Shah, S. A. A., and Bennamoun, M. (2018). *A Guide to Convolutional Neural Networks for Computer Vision*. Morgan & Claypool Publishers.

Kopittke, A. L. W. and Ramos, M. P. (2021). O que funciona e o que não funciona para reduzir homicídios no brasil: uma revisão sistemática. *Rev. Adm. Pub.*, 55(2):414–437.

Likforman-Sulem, L., Hanimyan, A., and Faure, C. (1995). A hough based algorithm for extracting text lines in handwritten documents. In *ICDAR*, volume 2, pages 774–777.

Louloudis, G., Gatos, B., Pratikakis, I., and Halatsis, C. (2009). Text line and word segmentation of handwritten documents. *Pattern recognition*, 42(12):3169–3183.

Lourenço, V., Mann, P., Guimaraes, A., Paes, A., and de Oliveira, D. (2018). Towards safer (smart) cities: Discovering urban crime patterns using logic-based relational machine learning. In *IJCNN*, pages 1–8. IEEE.

Mitchell, T. M. and Mitchell, T. M. (1997). *Machine learning*. McGraw-hill NY.

Nikitha, A., Geetha, J., and JayaLakshmi, D. (2020). Handwritten text recognition using deep learning. In *RTEICT*, pages 388–392.

Plamondon, R. and Srihari, S. (2000). Online and off-line handwriting recognition: a comprehensive survey. *IEEE TPAMI*, 22(1):63–84.

Purohit, A. and Chauhan, S. (2016). A literature survey on handwritten character recognition. *International Journal of Computer Science and Information Technology*, 7:1–5.

Reis, D., Melo, A., Coelho, A. L. V., and Furtado, V. (2006). Towards optimal police patrol routes with genetic algorithms. In *IEEE ISI*, pages 485–491.

SEPM (2015). Instrução normativa pmerj/emg-pm-3 nº 48 de 30 de dezembro de 2015.

Silva, G. (2021). Transcrição automática de textos em português escritos à mão usando deep learning. Bachelor's thesis, Universidade Federal Fluminense, Brasil.

Tomic, M., Grzunov, L., and Ivanovic, M. D. (2021). Crowdsourcing transcription of historical manuscripts: Citizen science as a force of revealing historical evidence from croatian glagolitic manuscripts. *Educ. Inf.*, 37(4):443–464.