Desenvolvimento de um algoritmo para detecção de caracteres: um estudo de caso para aplicação no Ministério Público Federal.

Ana Beatriz de Souza[1], Leila Cristina C. Bergamasco[2]
[1,2]Departamento de Ciência da Computação, Centro Universitário FEI
unifasouza@fei.edu.br leila.cristina@fei.edu.br

Resumo: Nos últimos anos, o uso da inteligência artificial (IA) em órgãos governamentais, como o Ministério Público Federal (MPF), se popularizou para aprimorar a atuação na proteção da sociedade, abordando questões como segurança digital e crimes cibernéticos. O MPF estabeleceu um portal para denúncias anônimas, no qual as evidências de imagens e vídeos são submetidas e analisadas manualmente. Este projeto de pesquisa visa utilizar inteligência artificial para melhorar a análise de denúncias feitas ao Ministério Público Federal. Isso será feito através da extração de texto de imagens enviadas como evidência, facilitando a identificação e classificação de casos, especialmente relacionados a crimes online.

1. Introdução

Um problema recorrente combatido pelo MPF é o discurso de ódio, o qual consiste em mensagens ofensivas e discriminatórias sobre determinada pessoa, população ou instituição. Com o advento da internet esses discursos passaram a ser disseminados em altíssima velocidade e a terem alcance global. A possibilidade de anonimato e a velocidade na disseminação das mensagens via internet, encorajam manifestações preconceituosas de todo tipo. No primeiro semestre de 2022 houve um aumento de 67,5% casos de discurso ódio comparado ao mesmo período de 2021, isso representa aproximadamente 24 mil casos (SAFERNET, 2022).

Uma das formas de detecção de tais crimes é por meio de denúncias no próprio site do Ministério, no qual o usuário pode enviar evidências por meio de imagens, vídeos e áudios que possam comprovar tal crime. Após o envio da denúncia, ela é analisada por uma equipe de especialistas que avalia se é válida e, caso positivo, como realizar a investigação e responsabilização dos indivíduos encarregados pelos crimes. Essa tarefa de análise é feita de forma manual, acarretando sobrecarga nos especialistas criminais e, consequentemente, um tempo maior de resposta.

Dessa forma, a aplicação de algoritmos de inteligência artificial (IA) que auxiliem na filtragem dessas denúncias pode otimizar o tempo gasto nessa etapa de pré-análise. Uma etapa anterior a essa classificação é detectar textos contidos em imagem e adequá-los para que possam ser processados posteriormente por tais algoritmos. Nesse sentido, o presente projeto de pesquisa tem por objetivo implementar e validar técnicas de Reconhecimento óptico de caracteres (*Optical Character Recognition*-OCR) em linguagem Python para, a partir de imagens, extrair textos e armazená-los de forma adequada para a futura utilização de algoritmos de mineração de texto.

Essa pesquisa é parte de um projeto maior que envolve a parceria do Centro Universitário FEI e o MPF de São Paulo - divisão de crimes cibernéticos, no qual se planeja implementar algoritmos que auxiliem na classificação de discursos de ódio e pornografia infantil.

2. Metodologia

O objetivo central deste projeto de pesquisa é a implementação e avaliação de técnicas de reconhecimento de caracteres, utilizando algoritmos de inteligência artificial. O projeto será conduzido através de seis etapas interligadas.

Primeiramente, será realizada uma pesquisa bibliográfica aprofundada, explorando as documentações relevantes. A formação da base de dados será uma fase inicial, com testes realizados em bases públicas já contendo textos de referência e saídas esperadas.

A etapa seguinte compreende o desenvolvimento de técnicas computacionais, empregando a pesquisa bibliográfica, especialmente sobre a biblioteca '*pytesseract*', para implementar o algoritmo e possíveis processos de pré-processamento em Python.

$$CER - \frac{S + D + I}{S + D + I + C}$$

$$WER - \frac{Sw + Dw + Iw}{Nw}$$

Figura 1 – Métricas CER e WER

Uma vez implementado, o método será minuciosamente avaliado e analisado, considerando as métricas de Taxa de Erro de Caractere (*Character Error Rate* - CER) e a Taxa de Erro da Palavra (*Word Error Rate* - WER).(Figura 1 e Figura 2).

A métrica CER é baseada na distância Levenshtein e ela representa o número mínimo de operações em nível de caractere necessárias para transformar o texto de referência na saída do OCR. Na primeira equação, *S* representa o número de erros por substituições, *D* significa o número de caracteres ausentes, *I* representa o número de inserções incorretas e *C* representa o número de caracteres corretos. A segunda fórmula, para o cálculo da WER, é a mesma que a da métrica CER, porém a WER opera em função da palavra e representa o número de substituições (*Sw*), exclusões (*Dw*) ou inserções (*Iw*) de palavras necessárias para transformar uma frase em outra e *Nw* representa o número total de caracteres.

Avaliações iniciais serão conduzidas utilizando bases de dados genéricas, seguidas por avaliações na base de dados real do MPF, proporcionando um panorama completo da eficácia do método.

Por fim, a documentação e divulgação dos resultados serão essenciais. Relatórios e artigos científicos serão elaborados com base nos resultados parciais e finais do projeto de pesquisa, com o intuito de compartilhar conhecimento e insights em congressos científicos e outros meios apropriados.

3. Conclusões

Para realizar o projeto, as seguintes etapas foram concluídas:

1. Formação da base de dados: realização de testes em bases de dados públicas, que já possuem textos de referências e respectivas saídas esperadas.
2. Desenvolvimento de técnicas computacionais: a partir da pesquisa bibliográfica realizada, principalmente no que tange o aprofundamento sobre a biblioteca *pytesseract*, o algoritmo e possíveis técnicas de pré-processamento foram implementados em Python.
3. Limpeza de dados: aplicação da biblioteca 'regex', a qual é especializada em identificar e manipular padrões de formatação nos dados. Sua utilização permitiu a remoção de elementos indesejados (ruídos).
4. Avaliação e Análise: o método desenvolvido será avaliado considerando as métricas descritas anteriormente. Foram feitas avaliações com bases de dados genéricas.

Na figura abaixo, foi utilizado uma postagem como exemplo para realizar o reconhecimento óptico de caracteres através da biblioteca *pytesseract*.

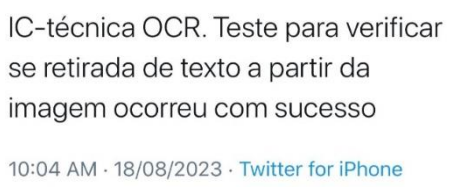

Figura 2 – Exemplo de publicação para a extração de texto

Percebe-se que, apesar do algoritmo conseguir extrair o texto a partir da imagem, ainda é necessário aperfeiçoá-lo para que sejam excluídos os espaços e as informações como data e hora da postagem. Tais ajustes foram possíveis com o auxílio das técnicas de mineração de texto exemplificado na figura abaixo:

Figura 3 – *Printscreen* do terminal com a Saída da OCR

É possível notar que após realizar esses ajustes com o texto obtido, se torna mais fácil a manipulação dos dados acerca do método avaliativo, uma vez que foram retirados os espaços adicionais e informações desnecessárias.

Ao conduzir a avaliação, identifica-se tal método adotado para análise: Inicialmente, o texto foi processado e dividido em duas listas distintas, uma contendo cada palavra extraída do texto e outra com cada caractere individual. A lista de palavras foi utilizada para calcular a métrica WER, comparando palavra por palavra com a referência. Por sua vez, a lista de caracteres foi empregada na métrica CER, realizando uma comparação minuciosa caractere por caractere. Após a constatação de que as listas de referência e de teste eram idênticas, a equação correspondente foi aplicada, resultando na obtenção dos valores esperados para ambas as métricas.

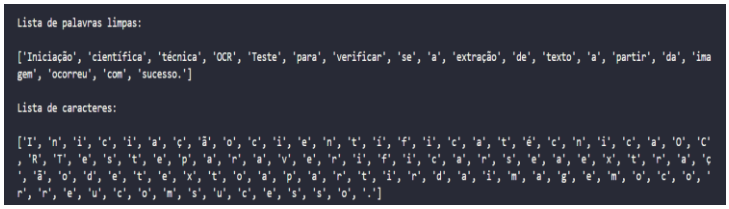

Figura 4 – *Printscreen* do terminal com as listas

Diante dos resultados obtidos e da análise das métricas de Word Error Rate (WER) e Character Error Rate (CER), pode-se afirmar que o método empregado neste estudo demonstrou eficácia na avaliação da precisão do reconhecimento de caracteres em textos OCR. Portanto, com base nos resultados deste estudo, conclui-se que o método apresenta um forte potencial para contribuir significativamente no avanço da precisão e confiabilidade do reconhecimento de texto por meio de tecnologias OCR. Essa pesquisa fornece uma base para investigações futuras e destaca a importância contínua de aprimoramentos no campo do reconhecimento óptico de caracteres.

5. Referências

[1] FERREIRA ALVES, Neide. Estratégias para melhoria do desempenho de ferramentas comerciais de reconhecimento óptico de caracteres. 2008. Diss. (Mestrado) – Universidade Federal de Pernambuco.
[2] LEUNG, Kenneth. Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER). [S.l.: s.n.], 2022.
[3] MEMON, Jamshed et al. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, IEEE, v. 8, p. 142642–142668, 2020.
[4] PYTHON, Real. Setting up a Simple OCR Server. [S.l.: s.n.], 2022.
[5] SAFERNET. Safernet aponta que discurso de ódio cresceu nas duas últimas eleições. [S.l.: s.n.], 2022.

Agradecimentos

À instituição FEI, seu corpo docente, direção e administração que oportunizaram a construção deste projeto.

[1] Aluno de IC do Centro Universitário FEI. Projeto com vigência de 04/2023 a 04/2024.