Detecção de Documentos Acadêmicos Falsificados: Uma Solução Baseada em Aprendizado de Máquina

Samuel M. Ransolin[1], **Giovana N. Inocêncio**[1], **Jean E. Martina**[1]

[1]Departamento de Informática e Estatística – Centro Tecnológico
Universidade Federal do Santa Catarina (UFSC) – Florianópolis, SC – Brasil

samuel.moreira.ransolin@grad.ufsc.br,
{giovana.inocencio,jean.martina}@ufsc.br

Abstract. *FIXME*

Resumo. *Nos últimos anos, no Brasil, o crescimento de ingressantes, de formandos e de instituições de ensino superior intensificou os desafios relacionados à validação de certificados acadêmicos, já que a verificação é majoritariamente manual, sujeita a erros e a aceitação de fraudes. Este trabalho revisita o estado-da-arte em detecção de documentos falsificados via aprendizado de máquina, e propõe um protótipo híbrido que combina análise multimodal, clustering, detecção de anomalias e classificação por grau de legitimidade. Ao integrar o protótipo à Jornada do Estudante, documentos podem ser validados automaticamente antes do registro em sua rede distribuída, aumentando a segurança e a confiabilidade do credenciamento.*

1. Problemática

Com o decorrer da última década, presencia-se no Brasil crescimento contínuo na emissão de diplomas de ensino superior, com cifras que chegam ao aumento de mais de 31% de formandos desde 2013 [**?**]. Embora isso revele um saldo extremamente positivo, também traz à tona desafios que precisam ser superados, entre eles e temática explorada neste trabalho, a melhoria nos processos de regulação, supervisão e avaliação dessas emissões por parte do Ministério da Educação do Brasil (MEC).

Atualmente, a gerência, armazenamento e emissão de documentos acadêmicos, como diplomas e históricos escolares, é responsabilidade da instituição de ensino que os emite [**?**]. Além disso, o processo, burocrático e não computadorizado, é suscetível a erros e até mesmo fraudes devido à ausência de transparência e redundância [**?**]. Assim, essa falta de modernização deixa brechas que são conhecidas e utilizadas por agentes mal-intencionados, possibilitando a criação de falsas instituições, especializadas na venda de pacotes que incluem diversos certificados contrafeitos amparados em documentos oficiais adulterados, de forma a conferir aparência de legalidade a diplomas sem qualquer base acadêmica real [**?**].

É neste cenário que o MEC, em parceria com o Ministério da Economia e diversas universidades federais, disponibiliza o sistema da Jornada do Estudante, que permite que discentes acompanhem suas trajetórias estudantis junto ao acesso a seus documentos acadêmicos pertinentes. Além disso, esse sistema também pode tornar-se uma plataforma conjunta para a emissão e registro destes certificados e até mesmo dados regulatórios das instituições de ensino superior [**?**]. Em acordo a essa iniciativa, o presente trabalho

trata da implementação e validação de um protótipo de software que combina diferentes técnicas de aprendizado de máquina, capaz de identificar certificados falsos antes de sua inserção neste ambiente.

2. Estado da Arte

A pesquisa acadêmica sobre identificação de documentos falsificados é escassa, especialmente quando comparada aos estudos sobre detecção de fraudes. Enquanto a detecção de fraudes foca em adulterações de arquivos originais (como mudança de notas, datas ou nomes), a identificação de documentos falsificados busca reconhecer aqueles completamente forjados desde sua criação, sem emissão por instituições oficiais. A área é predominantemente abordada através de técnicas de visão computacional, incluindo autoencoders convolucionais para análise de tintas em imagens hiperespectrais, métodos não supervisionados baseados em correlações espectrais, e reformulação do problema como comparação de grafos via OCR.

As abordagens mais robustas combinam múltiplas tecnologias para melhorar a detecção. O trabalho de integração entre blockchain e aprendizado de máquina para diplomas propõe um pipeline de quatro etapas: pré-processamento em nuvem, detecção de artefatos suspeitos com Faster R-CNN, segmentação pixel-level com Mask R-CNN, e consenso distribuído através de algoritmos de ML multicamadas. Paralelamente, pesquisas em análise multimodal demonstram a eficácia da fusão de características textuais e visuais, combinando extração semântica via ULMFiT e FastText com features visuais de redes convolucionais, alcançando 93,6% de acurácia na classificação de documentos.

Abordagens alternativas tratam a autenticação como problema de agrupamento, utilizando K-means para identificar padrões consistentes em documentos legítimos e classificando novos documentos pela proximidade aos centroides estabelecidos, atingindo 86,53% de acurácia.

3. Metodologia

Referências