Tamper Detection in Academic Credentials

Name: Yukti Juneja

Role: AI Development Intern

Date: May 2025

1. Introduction

Educational document forgery is a growing concern in both the academic and job markets, with a significant 15–20% of credentials having evidence of tampering. Most alterations involve altering grades, generating seals, or even metadata changes. This project seeks to create an automated detection system that can detect potential tampering in PDF educational documents such as degrees, transcripts, and professional certificates.

2. Objective

To create a prototype system that identifies indications of tampering based on:

- PDF metadata inconsistencies
- Text-level or visual content discrepancies
- Layout or formatting discrepancies

3. Methodology

There are two significant detection aspects:

A. Metadata Analysis (using PyPDF2)

- We pull embedded metadata such as creation date, modification date, author, and producer.

- If the modification date differs substantially from the creation date (with no apparent reason), it's an indication of a problem.

- HELPful in detecting if the document has been modified after issuance.

B. OCR-based Text Comparison (through Tesseract + pdf2image)

- PDFs are scanned into images, and then Optical Character Recognition (OCR) is used to extract text.

- Text extracted from an "original" and a "tampered" one is compared through Python's difflib.

- Word-level differences (e.g., grade modification or name change) are highlighted.

- Good for identifying changes in scanned documents or image tampering.

4. Detected Tampering Patterns

Document Type	Typical Tampering Signs
Degree Certificates	Altered degree type, date changes, name changes
Academic Transcripts	Grade alterations, courses added or deleted
Professional Certifications	Validity extension, changed skill level

5. Challenges & Trade-offs

Factor	Metadata Analysis	OCR-based Comparison
Speed	Fast	Moderate
Accuracy	High(for Metadata only)	High(if document is legible and unambiguous)
Robustness	Weak for scanned or image PDFs	Copes with images but relies on OCR quality
Complexity	Low	Medium

Challenge 1: Low scan quality decreases OCR accuracy.

Challenge 2: Some PDFs have metadata stripped from them, rendering that approach useless.

6. Future Scope

- In order to make the system production-capable:
- Implement OpenCV for matching templates and verification of seal/logo.
- Implement NLP models to scan for semantic inconsistency in document language.
- Implement blockchain to ensure authenticity at source.
- Train a classifier for identifying forged signatures or counterfeit stamps.

7. Conclusion

The prototype effectively detects generic manipulations in educational documents through metadata and OCR-based anomaly detection. Although useful for simple tampering, it can be extended through image processing and machine learning methods to detect more sophisticated forgeries. The system can be a starting point for an extended credential validation module employed by universities or companies.