National School of Arts and Crafts (ENSAM)

Rabat, Morocco

Technical Report

Multimodal Document Classification

Using NLP, Computer Vision, and OCR

Author

ZEMBA Wendemi Brice Romeo

bricezemba336@gmail.com

Academic Context

Personal Research Project

Artificial Intelligence and Data Science

Repository

`https://github.com/BriceZemba/Document-classification-NLP-CV`

January 2026

Abstract

This technical report documents the development of a sophisticated multimodal document classification framework specifically engineered for the automated processing of heterogeneous Curriculum Vitae (CV) datasets. Recognizing the limitations of unimodal systems in handling complex document layouts, this research proposes a hybrid pipeline that synergistically integrates Natural Language Processing (NLP), Computer Vision (CV), and Optical Character Recognition (OCR).

The architecture leverages a dual-stream feature extraction process: a text-based stream utilizing Transformer-based embeddings for semantic analysis, and a vision-based stream employing Convolutional Neural Networks (CNNs) to capture spatial and structural document features. By implementing a late-fusion strategy, the system effectively manages multimodal representations, significantly enhancing classification accuracy in the presence of scanned noise, varied formatting, and multilingual content.

A core focus of this work is the evaluation of model robustness and feature importance, addressing the "black-box" nature of deep learning through attention map visualizations. This project demonstrates the practical application of automated data cleaning and mapping for modernizing data-management infrastructures. The findings suggest that integrating structural layout information with textual context provides a more reliable and interpretable decision-making process for real-world document understanding applications, reducing manual error rates and optimizing system efficiency.

Keywords: Document Classification, NLP, Computer Vision, OCR, Multimodal Learning.

Contents

1 Introduction

The automation of document analysis has become essential in modern recruitment systems, digital archiving, and information management platforms. Curriculum vitae (CVs) are particularly challenging due to their heterogeneous layouts, multilingual nature, and frequent reliance on scanned documents.

This project proposes a multimodal classification pipeline that leverages both textual and visual information to improve robustness and classification accuracy.

2 Problem Statement

The goal of this work is to automatically classify CV documents into predefined categories while addressing the following challenges:

- Variability in document formats (PDF, images)
- Multilingual textual content
- Low-quality scans and OCR noise
- Layout-dependent semantic information

3 System Architecture

The system follows a modular architecture composed of five main stages:

- Document ingestion and conversion
- Text extraction and OCR
- Visual feature extraction
- Multimodal feature fusion
- Document classification

4 Data Processing Pipeline

4.1 Document Ingestion

Documents are accepted in PDF and image formats. PDF files are processed using `PyMuPDF` to extract both raw text and page images.

4.2 Optical Character Recognition

OCR is performed using `Tesseract` with multilingual support (French, Arabic, English), enabling text extraction from scanned documents.

4.3 Text Preprocessing

Extracted text undergoes normalization, tokenization, and vectorization before being passed to the NLP classification module.

5 Visual Feature Extraction

A convolutional neural network based on the ResNet architecture is employed to capture layout and structural patterns from document images. Residual connections allow deeper representations while maintaining training stability.

6 Multimodal Feature Fusion

Textual and visual feature vectors are concatenated into a unified representation. This fusion strategy enables complementary learning and improves classification robustness when one modality is degraded.

7 Model Training and Evaluation

The model is trained using supervised learning with cross-entropy loss. Performance is evaluated using accuracy, precision, recall, and F1-score.

8 Limitations

- Sensitivity to OCR errors in very low-resolution documents
- Increased computational cost due to multimodal processing
- Limited generalization to unseen document structures

9 Future Work

Future extensions include:

- Transformer-based text encoders
- Layout-aware multimodal models (LayoutLM, Donut)

- End-to-end document understanding architectures

10 Conclusion

This work demonstrates the relevance of multimodal approaches for document classification. By combining NLP, computer vision, and OCR, the system effectively handles real-world variability in CV documents and provides a strong foundation for intelligent document analysis systems.

Project Repository: `https://github.com/BriceZemba/Document-classification-NLP-CV`