LatIA. 2025; 3:315
doi: 10.62486/latia2025315
ISSN: 3046-403X

ORIGINAL

From Manual Review to AI Automation: an NLP-Powered System for Efficient CV Processing in Academic Admissions

De la revisión manual a la automatización mediante IA: un sistema basado en PNL para procesar eficazmente los CV en las admisiones académicas

Nadia Chafiq[1,2], Mohamed Ghazouani[3], Rokaya El Gounidi[1,2]

[1]Ecole Normale Supérieure- ENS -UH2C, Laboratoire Mathématiques, Intelligence Artificielle et Digital Learning, Casablanca, Morocco.
[2]Département Sciences de la Communication et Humanités, Faculté des Sciences Ben M'Sick-FSBM – UH2C, Casablanca, Morocco.
[3]Laboratoire Technologie de l'Information et Modélisation, Faculté des Sciences Ben M'Sick-FSBM – UH2C, Casablanca, Morocco.

Cite as: Chafiq N, Ghazouani M, El Gounidi R. From Manual Review to AI Automation: An NLP-Powered System for Efficient CV Processing in Academic Admissions. LatIA. 2025; 3:315. https://doi.org/10.62486/latia2025315

Submitted: 05-06-2024 **Revised:** 17-12-2024 **Accepted:** 23-05-2025 **Published:** 24-05-2025

Editor: Dr. Rubén González Vallejo

Corresponding author: Nadia Chafiq

ABSTRACT

Manual screening of thousands of admissions of master's program applications at Hassan II University of Casablanca is a time and labor-intensive task. Towards this challenge, we designed a machine-based solution utilizing Natural Language Processing (NLP) for summarization and CV ranking on a large set of CVs. Our solution relies on pre-trained spaCy and Hugging Face Transformers-based Named Entity Recognition (NER) models for the retrieval of information such as education, experience, and skills. We then incorporated extractive summarization by using BERT-based models for the selection of the most informative sentences and then the abstractive summarization by utilizing advanced language models such as LLAMA for the summaries to be coherent and easy. We verified our system by conducting a case study of the master's program of Big Data and Data Science by running a set of 2,325 CVs. The model gave very good results like a 72,67 % ROUGE-1 Recall, 74,32 % ROUGE-2 Recall, 73,15 % ROUGE-1 Precision, 57,28 % ROUGE-2 Precision, and 82 % Named Entity Recognition (NER) Precision. The system processed a CV on average in 3,84 seconds. We also integrated a conversation bot (chatbot) that allows admissions teams to search the CVs uploaded in real time for improved decision-making effectiveness and significantly decreasing the administrative burden. The promise of NLP-driven automation stands out from this research as a scalable as well as efficient method of screening numerous applicants.

Keywords: CV Summarization; NLP; Named Entity Recognition; Extractive Summarization; Abstractive Summarization; Candidate Ranking; Higher Education.

RESUMEN

La selección manual de miles de solicitudes de admisión a programas de máster en la Universidad Hassan II de Casablanca es una tarea que requiere mucho tiempo y trabajo. Para hacer frente a este reto, hemos diseñado una solución basada en máquinas que utiliza el Procesamiento del Lenguaje Natural (PLN) para resumir y clasificar un gran conjunto de currículos. Nuestra solución se basa en modelos preentrenados de reconocimiento de entidades con nombre (NER) basados en spaCy y Hugging Face Transformers para la recuperación de información como la formación, la experiencia y las habilidades. A continuación, incorporamos el resumen extractivo mediante modelos basados en BERT para la selección de las frases más informativas y, después, el resumen abstractivo utilizando modelos lingüísticos avanzados como LLAMA

para que los resúmenes sean coherentes y sencillos. Verificamos nuestro sistema realizando un estudio de caso del programa de máster de Big Data y Ciencia de Datos ejecutando un conjunto de 2.325 CVs. El modelo dio muy buenos resultados, como un 72,67 % de ROUGE-1 Recall, un 74,32 % de ROUGE-2 Recall, un 73,15 % de ROUGE-1 Precision, un 57,28 % de ROUGE-2 Precision y un 82 % de Named Entity Recognition (NER) Precision. El sistema procesó un CV en una media de 3,84 segundos. También integramos un bot de conversación (chatbot) que permite a los equipos de admisiones buscar en los CV cargados en tiempo real para mejorar la eficacia de la toma de decisiones y disminuir significativamente la carga administrativa. De esta investigación se desprende la promesa de la automatización basada en la PNL como método escalable y eficiente para seleccionar a numerosos solicitantes.

Palabras clave: Resumen CV; PLN; Reconocimiento de Entidades con Nombre; Resumen Extractivo; Resumen Abstractivo; Clasificación de Candidatos; Educación Superior.

INTRODUCTION

The Ministry of Digital Transition and Administration Reform, in cooperation with the Ministry of Higher Education, Scientific Research, and Innovation, and the Ministry of Economy and Finance, initiated a national program to fortify the digital talent pipeline of Morocco by 2027. The initiative sought to triple the number of Master's and Engineering Diploma graduates produced by the country's public universities each year from a current level of 8 000 up to 22 500 by the end of the program. To align higher education institutions with the needs of the labor market, 183 new accredited courses of study within key digital domains such as Artificial Intelligence (AI), Big Data, Cybersecurity, Data Science, Cloud Computing, and the Internet of Things (IoT) were introduced for the 2023–2024 academic year. While the initiative represents a significant step towards developing the digital workforce of Morocco, the effort also imposed operational challenges on the country's universities, the most significant of these relating to the handling of the influx of application submissions for the popular courses of study. Institutions were faced with unprecedented numbers of applications, often the thousands of applicants per course of study, putting a bottleneck on admissions. Each CV needs to be manually assessed and ranked by selection committees, a time-consuming task that stretches resources and lengthens the decision-making time. Legacy admissions procedures ask faculty and administrators to wade through vast application materials, check credentials, and evaluate candidates' fit for programs. For instance, one Master's program could get thousands of applications, taking weeks or even months of work to shortlist about 150 applicants for written tests. Only 25 get in the end per program. It not only lengthens time cycles but also imposes risks of inconsistency and subjectivism because evaluators use different criteria.

The increased administrative load has become a key challenge that draws away faculty time from fundamental tasks such as instruction and scholarship. The solution involves having scalable approaches that ensure fairness and rigor while optimizing workflows.[1,2,3]

Artificial Intelligence (AI),[4,5,6] and more specifically Natural Language Processing (NLP), hold transformative power when it comes to CV screening. The new NLP platforms like spaCy, Hugging Face Transformers, the BERT model, and LLAMA mean quick and accurate key candidate information extraction. Universities can integrate these tools and cut time and effort, as well as provide consistent data-driven assessments. Automating CV analysis facilitates the prioritization of qualitative evaluations over administrative routines for admissions committees while producing improved efficiency as well as transparency. Static summarization on its own is not adequate. For empowering decision-makers, our offer includes a conversational AI (chatbot) based on a tuned-up NLP model. The facility supports dynamic interaction within CV information and provides several advantageous benefits:

- Interactive Queries: evaluators can pose specific questions (e.g., "Who among the applicants have experience with Python and have published AI work?") and get instantaneous, accurate answers.
- Time Efficiency: the automated bot eliminates unnecessary searches, allowing quick retrieval of candidate information.
- Scalability: the system manages simultaneous inquiries at times of maximum loads, allocating loads appropriately.
- 24/7 Accessibility: decision-makers access information anytime, bypassing traditional office-hour constraints.
- Shared Features: retrieved information can be stored and shared among teams for collaborative decisions.
- Customization: questions align to program specifications, like emphasizing experience in research for academic programs or leadership experience for management programs.

By embedding this chatbot into admissions workflows, universities can transform a tedious process into an agile, equitable, and user-centric system. The assistant not only accelerates evaluations but also enriches decision-making with deeper insights into candidate qualifications.

Related works

Numerous studies have explored the automation of CV summarization using Natural Language Processing (NLP), drawing on advancements in information extraction, optical character recognition (OCR), and text summarization techniques. These studies provide valuable insights into effective methodologies, tools, and approaches for addressing similar challenges. By examining prior research, we can identify the best practices, understand the limitations of existing solutions, and pinpoint areas for innovation. Building on these foundations, our study aims to develop a solution that not only integrates but also enhances state-of-the-art NLP technologies to meet the specific demands of CV analysis and ranking in academic admissions.

In one study[7] the authors highlight the challenges faced by HR professionals in CV screening, such as application overload, lengthy processing times, and human biases in selection. These issues hinder recruitment efficiency and fairness, underscoring the need for an automated solution. The study proposes an automated CV ranking system, using a dataset of 223 CVs collected from various fields. The CVs, initially in PDF format, were converted into CSV for easier processing. The preprocessing phase included text normalization (lemmatization, stemming) and keyword extraction to improve text representation. Sentence embeddings were generated using S-BERT, and cosine similarity was used to measure relevance between CV content and job descriptions, assigning a relevance score to each CV. The system demonstrated high efficiency, processing each CV in just 0,233 seconds, and achieved 90 % accuracy in identifying relevant candidates. However, limitations such as potential biases in NLP models and computational costs remain. Future work aims to optimize algorithms for better contextual understanding and expand the dataset to improve model robustness.

Another study[8] addresses the limitations of OCR, particularly in handling handwritten and printed texts, where variability in writing styles and recognition errors hinder NLP performance. The authors propose an integrated pipeline combining OCR and NLP techniques to improve data extraction accuracy. The methodology involves segmenting lines using grayscale conversion, Sobel edge detection, Horizontal Projection Profile (HPP) calculation, and an A* path planning algorithm. Convolutional Neural Networks (CNNs) classify lines as handwritten or printed, followed by character recognition using the Tr-OCR model. Post-processing leverages NLP models for error correction, including space management and spell-checking using dynamic programming. Results show that Tr-OCR outperformed PP-OCR with a lower Character Error Rate (CER), and the classification model achieved 96 % accuracy. However, limitations include dependence on specific datasets and difficulty generalizing to various document formats. Future work aims to improve the model's ability to handle diverse writing styles and extend its application to broader contexts.

A third study[9] focuses on automating CV summarization to address the challenge of manually extracting key information from large volumes of textual data. The research compares various advanced models, including pre-trained models like BART, T5, and PEGASUS, as well as an LSTM model. A dataset of 75 anonymized CVs was created, emphasizing key sections such as experience, skills, education, and other relevant information. Data preprocessing involved removing special characters, converting text to lowercase, and tokenization. The pre-trained models were fine-tuned on the CV dataset and evaluated on open-source datasets like Xsum, CNN/Daily Mail, Amazon Fine Food Review, and News Summary. Performance was evaluated using the ROUGE metric, with BART-Large demonstrating strong performance by generating concise and relevant summaries with high ROUGE scores. Limitations include the focus on CVs, limiting generalization to other document types, and dependence on data quality input. Future work could explore by applying this approach to other document types and enhancing the models to better handle un-structured data.

In a fourth study,[10] researchers propose a system to automate the extraction of critical information from CVs, addressing challenges such as format diversity and high application volumes inherent in manual screening. The methodology employs a structured pipeline comprising data preprocessing, segmentation, information ex-traction, and machine learning-based classification. PDF CVs are first converted to HTML using PDFminer, with relevant sections (e.g., education, work experience) extracted via BeautifulSoup. Segmentation identifies distinct CV sections by analyzing font size, style, and syntax trees, while information extraction combines Named Entity Recognition (NER) through NLTK and regex-based pattern matching to structure outputs into JSON format. For classification, Decision Tree (ID3) and Logistic Regression algorithms evaluate attributes like CGPA, skills, and achievements. Experimental results demonstrate 80–85 % precision with ID3 on a sample of 50 CVs, though Logistic Regression outperforms ID3 on larger datasets, highlighting scalability advantages. However, limitations persist, including difficulties with highly varied layouts and the need to discretize continuous attributes for ID3 compatibility. To address these gaps, future work proposes expanding the dataset, integrating advanced NLP models like BERT or GPT for nuanced contextual understanding, and adopting ensemble methods to enhance robustness and accuracy.

A fifth study[11] tackles the challenge of efficient CV processing in software engineering recruitment, leveraging NLP, character positioning, and regular expressions (regex) to enhance information extraction. The methodology combines spaCy for Named Entity Recognition (NER) to extract names, email addresses, phone numbers, degrees, and skills, along with word and phrase matching to identify technical competencies. A character positioning method detects key sections such as education or work experience. Regular expressions extract structured information like email addresses and phone numbers. Data was collected via web scraping of CVs and job descriptions from Merojob, with skills categorized into areas like Backend, Front-end, Mobile Development, and Machine Learning. Results showed an overall extraction success rate of 33,59 %, with better performance on structured CVs. However, the system struggled with unstructured CVs and identifying names and experiences. Limitations include a small dataset, difficulties with varied layouts, and dependence on predefined skills. Future improvements involve expanding datasets, integrating advanced NLP models, and adding OCR capabilities.

A sixth study[12] proposes a hybrid architecture combining BERT for extractive summarization and LSTM for abstractive summarization to enhance text summarization quality. The model uses BERT to select key phrases and LSTM to reformulate and paraphrase the extracted information. The model is optimized using Particle Swarm Optimization (PSO). The BBC News Summary dataset, containing 2 225 articles, was used for evaluation. Preprocessing included segmentation, tokenization, stop word removal, stemming, and lemmatization. Performance was evaluated using ROUGE scores, with the hybrid BERT-LSTM model outperforming existing approaches. Limitations include a correct extraction rate of 33,59 % and difficulties processing texts that deviate from trained structures. Future recommendations involve refining the model and integrating advanced architecture.

A seventh study[13] introduces a sophisticated chatbot system integrating NLP and AI techniques to optimize dialogue management and user satisfaction. The chat-bot architecture includes an interactive User Interface (UI), an NLP engine for tokenization and NER, an AI decision module, a response generation module, and data storage. The system achieved 92 % response accuracy, an average user satisfaction rating of 4,3 out of 5, and 88 % context retention in multi-turn dialogues. Limitations include an initial extraction rate of 33,59 % and challenges with unstructured inputs. Future improvements include multilingual support and API integrations.

An eighth study[14] proposes an automated system for CV screening using NLP and AI. The system processes PDF CVs perform preprocessing steps like morphological analysis and spellchecking and use syntactic and semantic analysis to rank candidates. The system processed 203,621 tokens and completed data processing in 23 seconds. Limitations include dependence on CV format and quality. Future improvements involve enhancing format support and extending the system to other recruitment sectors.

Finally, a ninth study[15] leverages advanced Large Language Models (LLMs) like GPT-3.5 and GPT-4 for CV classification, summarization, and evaluation. The framework uses a dataset of 1000 IT sector CVs, with information categorized into seven types. The system demonstrated efficiency, being 11 times faster than traditional methods, and achieved an F1 score of 87,73 % for CV classification. Limitations include potential biases in LLM training data and dependence on JSON standardization. Future enhancements include testing on diverse sectors and integrating OCR techniques.

Table 1. Summarizes the literature survey

Paper	Objective	Techniques
[7]	Automate CV sorting for recruiters, reducing human bias and processing time.	Pretreatment data (lemmatization, keyword extraction), generation of embeddings with S-BERT, calculation of cosine similarity between CVs and job descriptions.
[8]	Improve optical character recognition (OCR) for handwritten and printed texts.	Segmentation with A* algorithm, classification with CNN, character recognition with Tr-OCR, post-processing with NLP (spelling correction and space management).
[9]	Automatically summarize CVs using advanced templates to improve efficiency and reduce human bias.	BART, T5, PEGASUS and LSTM models tested for abstract summarization. Pre-processing (tokenization, sup-pression of special features). Training with datasets such as CNN/Daily Mail and Xsum.
[10]	Efficiently extract and analyze CV information for automatic classification.	Extraction with NER (NLTK), segmentation via syntax and font styles, classification with ID3 and logistic regression.
[11]	Automate the extraction of key information from CVs for software engineering positions.	Use of spa-Cy for NER, regex to extract structured information, skill categorization with pre-defined data.

Our solution expands on and overcomes the shortcomings of previous work within this area as it brings forth a CV summarization and ranking system fueled by Natural Language Processing (NLP). Previous methods are surpassed by our system as it includes a conversation-based assistant (chatbot) that makes managers interactive by allowing them to directly engage with CVs that have been uploaded. The optimized NLP model utilized by the chatbot extracts structured data from documents and provides a more intuitive and intelligent means of pinpointing suitable applicants as opposed to keyword-based search methods or traditional static summarization.

METHOD

Our approach addresses a number of the most significant challenges facing CV processing, aiming both to overcome present obstacles and to create new opportunities. One of the primary difficulties lies in the diversity of CV submission formats, ranging from structured documents such as Word or PDF files to unstructured formats including scanned images and scanned PDFs. This variability complicates the extraction and organization of relevant information. Our approach responds to this challenge by implementing a robust and scalable system capable of supporting a wide range of input formats while ensuring accurate and efficient information extraction.

Data Collection and Preparation

Before building an automated CV processing system, it is essential to prepare and structure the dataset. This involves collecting a diverse range of CVs and ensuring they are in a format suitable for extraction and analysis.

To evaluate our system, we used a dataset of 2,325 CVs in PDF format. These CVs follow a structured layout, making them well-suited for initial model testing. However, in real-world scenarios, CVs come in various formats, including DOCX, plain text, scanned PDFs, and images. As a result, the system must be designed to handle these diverse formats effectively.

Since raw text extraction is a critical first step for any further processing, we use specialized tools to convert various document formats into structured text:

- Docling:[16] this tool extracts raw text from PDF, DOCX, PPTX, XLSX, HTML, and Markdown documents. It provides advanced capabilities for understanding page layouts, reading order, and table structures, ensuring that the extracted content maintains its original meaning.
- Optical Character Recognition (OCR): For scanned documents and images, we use Tesseract OCR to retrieve text from non-editable formats.

Once the text is extracted, a preprocessing step is applied to clean the data. This involves removing unnecessary characters, headers, and footers while preserving essential content. This step ensures consistency and readability before proceeding to information extraction.

Named Entity Recognition (NER) for Information Extraction

Extracting key information from CVs relies on identifying specific entities, such as names, job titles, work experience, education, and skills. This step is crucial for organizing a CV in a meaningful way and making the information usable by admissions evaluators.

We use pre-trained spaCy Named Entity Recognition (NER) models to detect standard entities like dates, locations, job titles, and organizations. However, generic NER models often struggle with domain-specific terminology, especially in technical fields like IT. To address this gap, we use an integrated approach that combines:

- Dictionary-based matching to detect IT-specific terms (e.g., programming languages, frameworks, and tools).
- Fuzzy Matching (using FuzzyWuzzy or RapidFuzz) to recognize skills that may be spelled differently across various CVs.
- Transformer-based models (e.g., Hugging Face Transformers) to capture more complex entities that traditional rule-based methods might overlook.

By blending these techniques, we significantly improve the accuracy of information extraction, ensuring that both skills and experience are reliably identified.

CV Summarization

After extracting key entities (e.g., names, job titles, educational background, and technical skills), the subsequent phase involves generating concise summaries that preserve essential information. These summaries facilitate a more efficient review process for Admissions Evaluators, enabling them to assess a candidate's profile without consulting the full document.

This study employs a hybrid technique that combines extractive and abstractive summarization methods:

- Extractive Summarization leverages approaches such as TF-IDF and BERT-based models to identify and retain the most relevant sentences from the original CV.
- Abstractive Summarization utilizes advanced natural language generation architectures (e.g., T5, BART, or GPT-based transformers) to restate critical details in a more cohesive, distilled format.

To further enhance clarity and navigability, the summarized content is organized into predefined categories, most commonly Skills, Work Experience, and Education. This structured presentation ensures that salient features of the candidate's background are readily accessible and easily interpretable. By merging extractive

and abstractive summarization techniques, our dual-strategy framework preserves the integrity of critical information while enhancing the clarity and conciseness of CV content. This hybrid methodology addresses a key challenge in academic admissions workflows, where time-sensitive evaluation and accuracy-driven candidate assessments are critical priorities. The integration of these complementary approaches ensures that evaluators receive both precise, data-rich insights and contextually coherent summaries enabling faster, more informed decision-making without compromising depth or reliability.

Job Fit Evaluation Using Cosine Similarity

In addition to summarizing CVs, our system assesses the alignment between a candidate's profile and specific job requirements through a semantic matching framework. This is accomplished by computing cosine similarity between vector representations of the candidate's skills, extracted via NLP, and the competencies outlined in the job description.

The methodology involves three key phases:

1. Embedding Generation: job descriptions are encoded into dense vector embeddings using Sentence-BERT (SBERT), a pre-trained language model adept at capturing semantic meaning.

2. Similarity Computation: cosine similarity metrics quantify the degree of alignment between candidate skills and job requirements, ensuring a data-driven evaluation of relevance.

3. Candidate Ranking: profiles are systematically ranked by similarity scores, enabling evaluators to prioritize candidates whose qualifications most closely match institutional criteria.

By automating the alignment assessment, this approach streamlines the shortlisting process, reducing manual screening time while enhancing objectivity. Evaluators can thus dedicate greater attention to nuanced qualitative assessments, ensuring that high-potential applicants are identified efficiently and equitably.

Dynamic Summarization

To address diverse recruitment needs, our system incorporates a modular summarization framework, enabling users to tailor the depth and granularity of generated summaries. This flexibility is achieved through two distinct operational modes:

1. Section-Based Summarization: focused on specific segments of a CV (e.g., professional experience, technical skills), this mode generates succinct, targeted overviews that highlight domain-relevant qualifications.

2. Full Summarization: this mode synthesizes the entire CV into a cohesive yet concise narrative, preserving critical details while eliminating redundancies.

By supporting these configurable modes, the system adapts to heterogeneous evaluator requirements from rapid competency screening to in-depth professional background analysis. Such adaptability not only streamlines the evaluation process but also enhances precision in identifying candidates whose profiles align with institutional or role-specific criteria.

Performance Evaluation and Measurement

Evaluating the efficacy of our system is imperative to validate its reliability and accuracy. To achieve this, we employ a rigorous evaluation framework comprising multiple performance metrics:

- Information Extraction Accuracy is quantified using Precision, Recall, and F1 scores, which collectively measure the system's ability to identify and categorize key entities such as skills, education, and professional experience.
- Summary Quality is assessed through ROUGE and BLEU metrics, enabling a systematic comparison between system-generated summaries and human-authored reference summaries.
- Processing Speed is monitored to determine the average time required to analyze a single CV, ensuring the system's scalability for high-volume admissions cycles.
- User Satisfaction is evaluated via structured surveys and qualitative feedback from admissions evaluators, capturing insights into usability and practical utility.
- Finally, Cross-Domain Adaptability is validated by testing the system on CVs spanning diverse sectors (e.g., IT, finance, healthcare), thereby demonstrating its robustness across heterogeneous datasets.

RESULTS

To further contextualize our results, we benchmark our approach against state-of-the-art AI-based CV summarization models, including GPT-4 and LLaMA-3. This comparative analysis not only highlights our system's competitive performance but also identifies opportunities for refinement, ensuring continuous improvement in accuracy, efficiency, and user-centric design.

By combining Optical Character Recognition (OCR), Named Entity Recognition (NER), Summarization, and Job Fit Evaluation, we offer a robust solution for processing CVs at scale. This blend of rule-based, machine-learning, and deep-learning techniques ensures a high level of flexibility and accuracy, making our system particularly

valuable to Admissions Evaluators seeking to optimize candidate screening. To clarify the inner workings of our solution, we have developed a System Architecture Diagram (figure 1). This illustration outlines the core components and their interactions from data ingestion and preprocessing through information extraction, summarization, and final presentation of results. It demonstrates how diverse file formats (including PDFs, DOCX files, and scanned images) are processed by text extraction modules such as Docling and Tesseract OCR, followed by a robust NLP pipeline that leverages spaCy, fuzzy matching, and Transformer-based models for NER. The diagram also showcases our Job Fit Evaluation module, which applies Sentence-BERT and cosine similarity to rank candidates according to their suitability. By presenting the system's structure visually, we underscore the modularity, scalability, and guiding design principles that shape our automated CV processing approach.

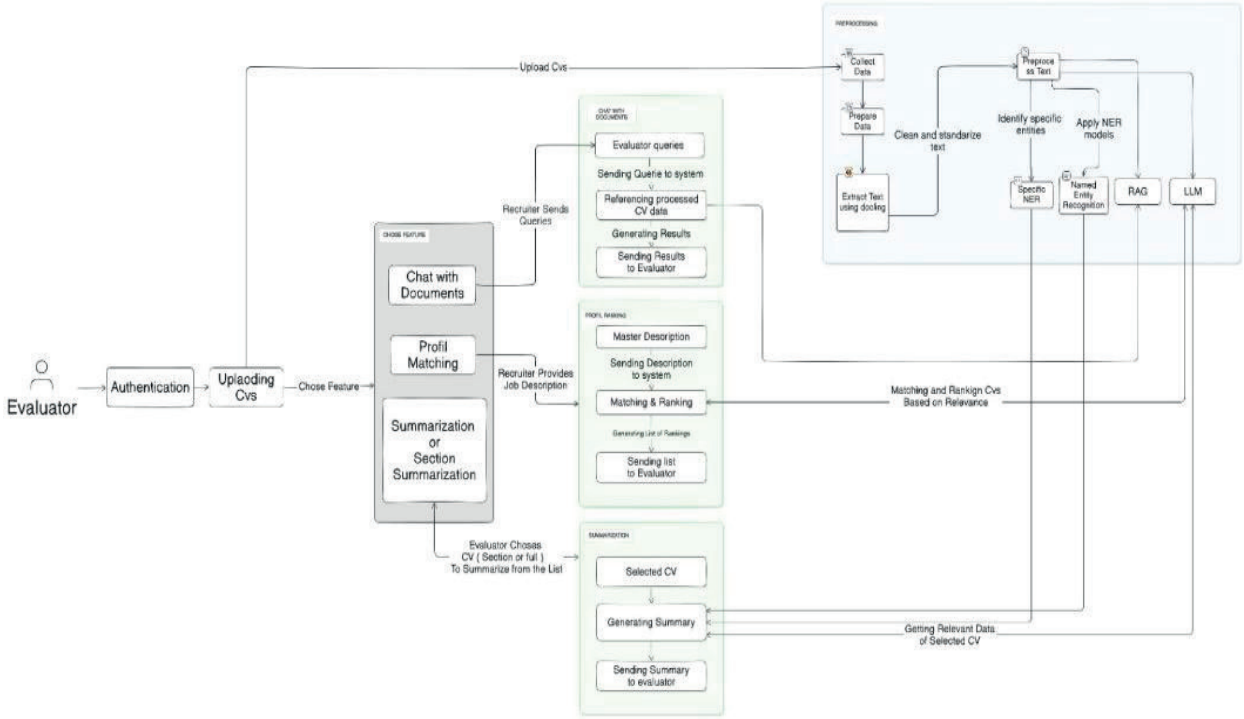

Figure 1. Overview of the Proposed System Architecture

Web application development (django framework)

To enhance system accessibility and user-friendliness, we implemented a web application using the Django framework. This interface allows users to upload their CVs and receive a structured summary in real time.

Key features include:

- Frontend for CV Upload: Users can submit CVs in PDF, DOCX, or image formats.
- Automated Backend Pipeline: The server-side logic handles Optical Character Recognition (OCR), Named Entity Recognition (NER), and summarization processes.
- Real-Time Output: The platform immediately displays summarized CVs and enables quick downloads.

For scanned PDFs and image-based inputs, Tesseract OCR is employed to convert content into text prior to subsequent processing. Final summaries are available in PDF or Word format for user convenience. To ensure smooth and intuitive user experience, we designed both a User Flow Diagram (UFD) and an Application Flow Diagram (AFD). The UFD (figure 2) illustrates the user's journey from accessing the system and uploading a CV to viewing and retrieving the processed summary. This diagram highlights the application's user-centric perspective, aligning its interface and functionality with typical workflows and user expectations.

Building upon the insights provided by the User Flow Diagram, the Application Flow Diagram (AFD, figure 3) presents an internal view of the system's processes and their interactions. This includes data ingestion, Named Entity Recognition (NER), summarization algorithms, and the integration of OCR for handling scanned files. By designing these diagrams, we aimed to optimize both the front-end usability and the back-end efficiency of our application, ultimately ensuring a seamless and effective experience for those overseeing candidate selection.

Upon successful login, users encounter a streamlined interface that facilitates batch uploading of multiple CVs in various formats (PDF, DOCX, or scanned images). Once the upload is complete, a comprehensive set of tools becomes available, enabling in-depth analysis and management of the submitted applications. As shown

in figure 4, the home page serves as an intuitive entry point for this process, guiding users seamlessly through CV submission and subsequent feature exploration.

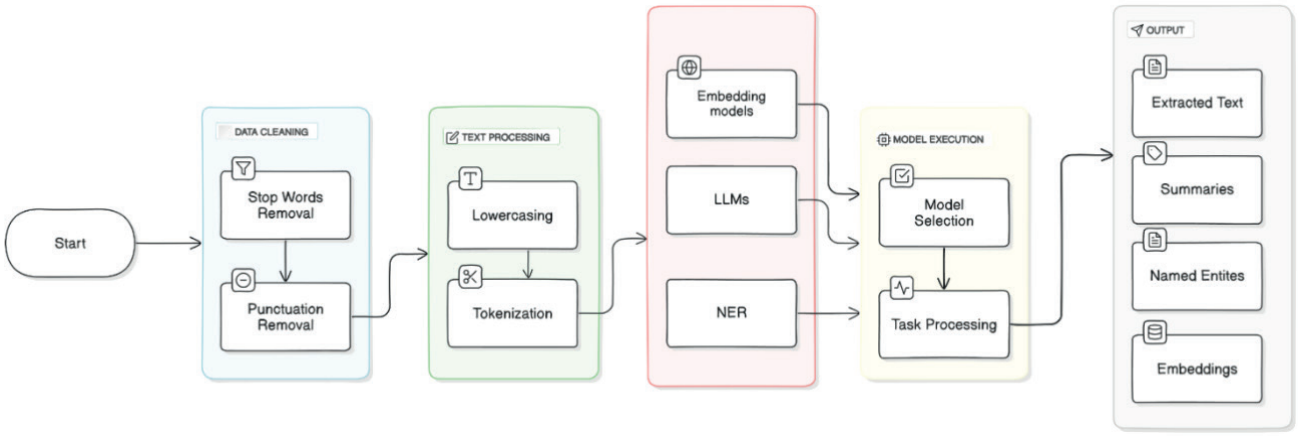

Figure 2. User Flow Diagram

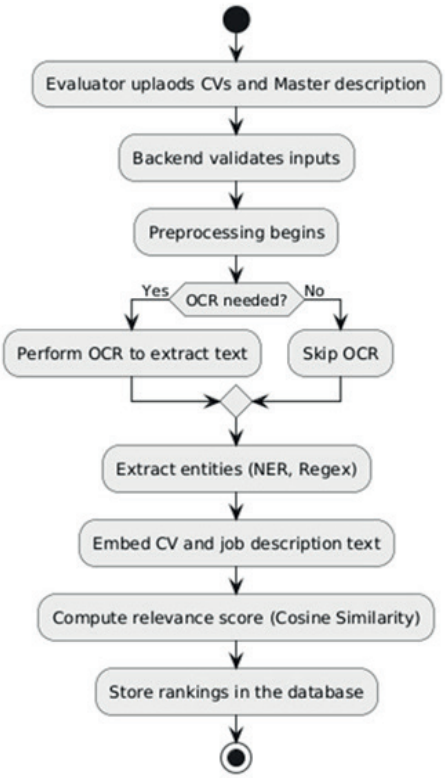

Figure 3. Application Flow Diagram

Figure 4. Web Application Home Page

As shown in figure 5, the CV summarization feature allows users to generate either a comprehensive overview of the uploaded CVs or a focused summary of specific sections (e.g., experience, skills, education).

Upon selection, an advanced NLP model (LLAMA3) creates the requested summary in near real-time. These summaries appear directly within the user interface, providing a concise, easily digestible representation of each candidate's profile.

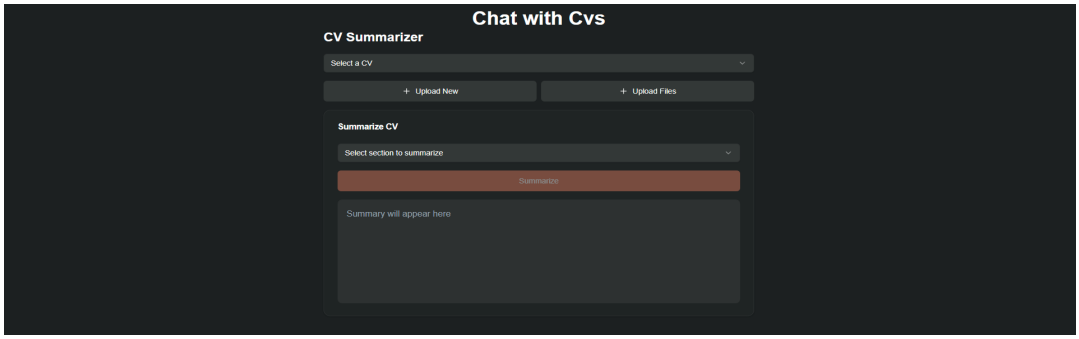

Figure 5. User Interface for Generating CV Summaries

As illustrated in figure 6, the CV ranking module begins with the evaluator describing the master's program, either by manual input or by importing a pre-existing description document. The system then automatically ranks the uploaded CVs based on their alignment with the specified master's program criteria, leveraging an integrated NLP-powered matching model. The resulting rankings are presented through an interactive interface, complete with filtering options to facilitate refined searches and the identification of the most suitable candidates for admission.

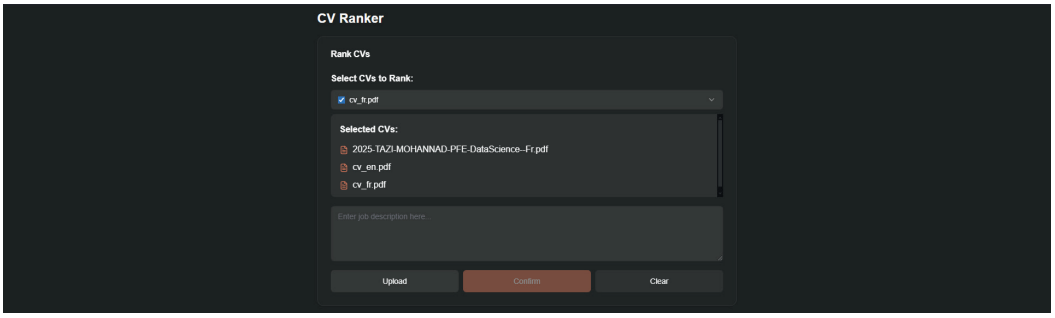

Figure 6. CV Ranking Module User Interface

As shown in figure 7, a conversational assistant enables evaluators to pose questions about the uploaded CVs. For example, an evaluator might ask, "Which candidates have a strong background in statistical analysis and machine learning, supported by relevant coursework or project experience?" The AI system then analyzes the CVs and provides a focused, data-driven response. This chatbot is powered by an optimized NLP model that extracts structured information from documents, enabling more effective candidate assessment.

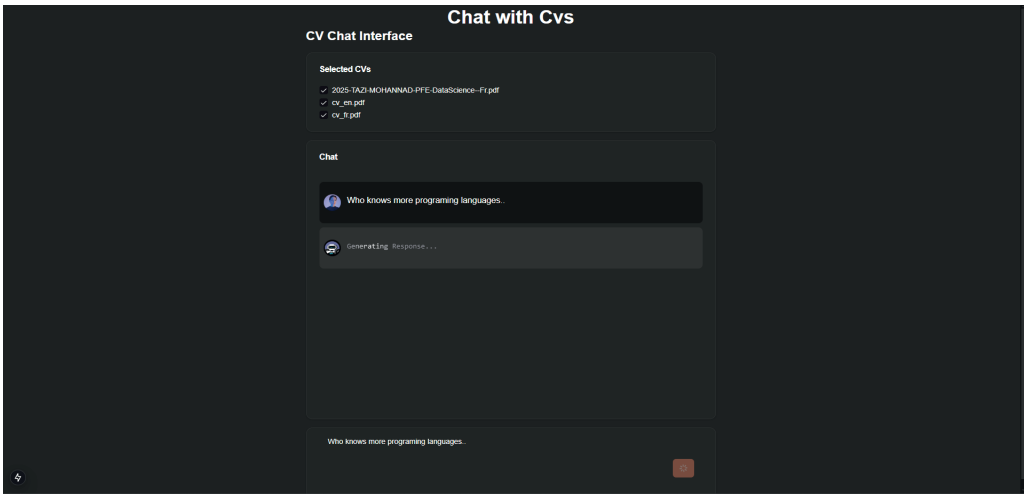

Figure 7. Chatbot Module for Questioning and Analyzing CV Data

DISCUSSION

Our methodology offers a holistic, scalable model for automated CV processing to deal with the heterogeneity in document formats and increasing candidate evaluation needs in the academic domain. Our system combines OCR, NLP tools, and semantic similarity computation and can achieve impressive results in extracting structured information and summarizing meaningful content. A major strength of our framework is its comprehensive support for multiple formats. CVs submitted in the real world regularly come in different shapes and forms and can be anything from DOCX files to scanned PDFs, and scanned images; this presents a significant challenge for standard parsers. We manage to consider this diversity of documents by coupling Docling with Tesseract OCR, achieving generalizability in practice. But on top of that, though it's still early, it's not clear whether even further fine-tuning and error handling would have been necessary to make out documents like handwritten or low-res scans consistently (you can look at red underlines not as the source of true stuff but as a close approximation for where the model felt less sure; it's appropriate to be cautious and correct number or kind of error that occur).

Another significant accomplishment is the application of the strategy of hybrid information extraction. The state-of-the-art generic NER models generally fail in specific domains such as engineering or computer science. To this end, we combined dictionary-based detection and fuzzy string matching with transformer-based models, achieving a drastic escalation in the ability to recognise technical skills and nuanced experiences. This multilayer architecture increases recall without decreasing precision, however, constant re-making of domain-specific dictionaries and model refinement is required as new lingo develops.

Our approach is characterized by its summarization component as well. By blending extractive and abstractive strategies, we achieve a compromise between factuality and linguistic consistency. The ability to produce either section or full summaries is well suited to different user preferences from the evaluator, screening rapidly or drilling down into the profiles. However, abstractive models also generated generalization errors, and therein, the need to combine feedback loops or human-in-the-loop checkers into abstractive models in order to improve the outputs. Our Job Fit Evaluation module based on Sentence-BERT and cosine similarity performed very well to sort the applications as per the match to the masters' program. This automated scoring facilitates objective judgment, reduces bias, and eases the workload of raters. In the future, one could consider using contextual embeddings or knowledge graphs to model more of the semantic match between applicant profiles and program goals. The web-based platform built with Django has maximized accessibility and real-time capability, connecting technical performance with user experience. "Pilot users are very happy with the ease of use and functionality; they have asked for deeper and more specific capabilities, and they've also reported higher levels of job satisfaction," says the IT director. However, wider testing for usability and across institutions and evaluator types will be required to assess the platform's generalizability and any potential adoption obstacles. Finally, the system's scalability has been confirmed by performance assessment (F1 scores since extraction and ROUGE/BLEU for summarization, computing speed, etc.). Cross-domain CVs from the finance and healthcare sectors show general capability, while small degradation in NER performance in non-technical domains suggests the necessity of domain-sensitive customization.

In conclusion, this work establishes strong grounds for intelligent CV processing in academic admissions by a merging of technical rigour and practical usability. Although there remain drawbacks, e.g., OCR quality differences, abstractive summarization drift, and evolving terminology, the modular and extensible architecture of our framework allows for painless integration of future revisions and improvements. Future work will refine generation-specific models, taking feedback from deployment in the wild and investigating explainable AI approaches that can increase evaluator confidence and understanding in automated suggestions.

CONCLUSIONS

Our automated CV summarization and ranking system represents a transformative approach to addressing the challenges of large-scale candidate evaluation. By harnessing advanced Natural Language Processing (NLP) and Artificial Intelligence (AI) technologies, we propose a robust and adaptable framework for processing CVs. The system's hybrid methodology, which integrates extractive and abstractive summarization techniques alongside a conversational assistant for interactive data exploration, establishes a new benchmark in automated candidate assessment. A central innovation of our system is the seamless integration of sophisticated NLP methods with an intuitive user interface, featuring dynamic summarization capabilities and an interactive chatbot. This design empowers universities to optimize their admissions workflows, reduce subjective biases, and more effectively identify top-tier applicants for their master's programs.

Our findings underscore the substantial benefits of NLP-driven automation, including a significant reduction in administrative workload and the establishment of a more consistent, data-informed shortlisting process. Looking ahead, we aim to refine the abstractive summarization component and expand the system's applicability to a broader spectrum of academic disciplines. Furthermore, as video CVs gain traction in recruitment processes, future research will explore the integration of video processing capabilities. This

includes leveraging speech-to-text technology for automated transcription, employing facial recognition for demographic insights, and utilizing sentiment analysis to evaluate communication styles. By incorporating these advanced features, the system will offer a richer, more nuanced assessment of candidate qualifications, further enhancing the impact of our automated CV processing solution.

REFERENCES

1. Yatsun O. Smart city concept: Integrating technology into municipal governance. Gentrification. 2025;3:85. https://doi.org/10.62486/gen202585

2. Dashdamirli R, Abdullayev V. Artificial intelligence-based smart city ecosystem development. Land and Architecture. 2025;4:180.

3. Kiran Vege H, Yandamuri SK, Vennela J, Venkat S. Ai for autonomous health care on diabetes diagnostics. South Health and Policy. 2025; 4:236. https://doi.org/10.56294/shp2025236

4. Lala G, Vugar A. Application of IoT and Sensor Technologies in Environmental Monitoring. Environmental Research and Ecotoxicity. 2025;4:170.

5. Alqaraleh M, Salem Alzboon M, Subhi Al-Batah M. Real-Time UAV Recognition Through Advanced Machine Learning for Enhanced Military Surveillance. Gamification and Augmented Reality. 2025;3:63. https://doi.org/10.56294/gr202563

6. Sousa L, Carvalho ML, Mestre R, Tomás J, Severino S, José H. Artificial Intelligence in Nursing: applications, challenges and future directions. Gamification and Augmented Reality. 2025;3:113.

7. Enhanced Resume Screening for Smart Hiring using Sentence-Bidirectional Encoder Representations from Transformers (S-BERT). 2024.

8. A Novel Pipeline for Improving Optical Character Recognition through Post-processing Using Natural Language Processing. 2023.

9. Abstractive Text Summarization for Resumes with Cutting Edge NLP Transformers and LSTM. 2023.

10. Analyzing CV/Resume Using Natural Language Processing and Machine Learning. 2022.

11. Automatic Software Engineering Position Resume Screening Using Natural Language Processing, Word Matching, Character Positioning, and Regex. 2022.

12. NLP-Based Automatic Summarization Using Bidirectional Encoder Representations from Transformers-Long Short-Term Memory Hybrid Model: Enhancing Text Compression. 2024.

13. Leveraging NLP and AI for Advanced Chatbot Automation in Mobile and Web Applications. 2021.

14. Resume Ranking Using Natural Language Processing. 2024.

15. Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening. 2024.

16. Auer C, Lysak M, Nassar A, Dolfi M, Livathinos N, Vagenas P, et al. Docling Technical Report. arXiv [Preprint]. 2024 Aug 18; https://arxiv.org/abs/2408.09869

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Nadia Chafiq, Mohamed Ghazouani,Rokaya El Gounidi.
Data curation: Mohamed Ghazouani,Rokaya El Gounidi.