

Development of a Web-Based Resume-Job Matching System Using Transformer Embeddings and Weighted Jaccard Similarity at Systems Plus College Foundation, Inc.

Kenneth C. Carillo ¹, Kurt Russel Cano ², Benedict C. Garcia ³, Clarence Paul Q. Lazaro ⁴

College of Computing and Information Sciences, Systems Plus College Foundation, Angeles City, Philippines

Private, Non-Sectarian Higher Education Institution , Philippines

**Corresponding Author Email: carillo.kenneth24@gmail.com*

ABSTRACT

Recruitment workflows often suffer from operational bottlenecks due to manual resume screening and rigid keyword matching, which fail to capture contextual qualifications and encounter extraction errors from diverse document layouts. To address these challenges, this study proposes the development of a web-based resume-job matching system for Systems Plus College Foundation, Inc. utilizing layout-aware visual Optical Character Recognition (OCR), Transformer-based embeddings, and Weighted Jaccard Similarity. The system integrates a visual OCR engine to extract structured text from complex, multi-column resume layouts and generates high-dimensional vector representations to measure contextual semantic alignment against job descriptions. It also utilizes Weighted Jaccard Similarity to evaluate explicit skill overlap based on term criticality, producing a calibrated candidate compatibility score. The proposed hybrid Transformer–Jaccard approach will be benchmarked against conventional keyword-based screening across four primary performance measures: layout-parsing accuracy, semantic extraction precision, runtime processing efficiency, and false rejection rates. Following a Developmental Research Design with a Quantitative Approach implemented through the V-Model methodology, the system offloads heavy processing tasks to asynchronous background queues while delivering ranked candidate analytics to a web dashboard. The study will also evaluate user acceptance based on ISO/IEC 25010 standards to confirm functional suitability and usability. The system is expected to improve initial screening accuracy over traditional keyword methods and provide human resource personnel with a transparent, data-driven decision-support tool for optimizing institutional hiring workflows.

Keywords: *Visual Optical Character Recognition, Transformer-based embeddings, Weighted Jaccard Similarity, Semantic alignment, Hybrid Transformer–Jaccard approach, V-Model methodology, ISO/IEC 25010 standards*

SECTION I — INTRODUCTION (THE PROBLEM AND ITS SETTING)

1.1 Introduction

Recruitment requires efficient resume screening to evaluate candidate qualifications accurately. Manual resume review and keyword searches create significant bottlenecks during high-volume hiring. Recruitment friction directly increases applicant drop-out. Bullhorn (2022) found that 75% of Generation Z candidates abandoned job applications due to slow communication, lengthy timelines, and complex screening. However, because this survey examined 1,000 Generation Z applicants in North America, these findings primarily reflect that specific demographic. Layout diversity creates another key technical barrier. Multi-column structures, tables, and graphics disrupt reading order and data extraction. Werner and Laber (2024) identified reading order reconstruction and section extraction as major practical hurdles in automated resume parsing. Furthermore, traditional keyword matching misses contextual semantics. Transformer-based models resolve this by

generating dense vector representations that capture underlying textual meaning beyond exact word matches. To address these challenges, this study proposes a Web-Based Resume-Job Matching System Using Transformer Embeddings and Weighted Jaccard Similarity at Systems Plus College Foundation, Inc. The system integrates document parsing, visual OCR, Transformer embeddings, and Weighted Jaccard Similarity to extract candidate data and evaluate job compatibility accurately.

1.2 Background of the Study

Automated recruitment systems simplify applicant tracking, but their accuracy depends on precise data extraction and job requirement comparison.

Diverse file formats and visual structures complicate automated parsing. Werner and Laber (2024) developed a resume parser to resolve reading order reconstruction and segment key sections like education and employment history. Incorporating Optical Character Recognition (OCR) and layout-aware processing preserves textual positioning in scanned documents. Zhu et al. (2025) proposed a layout-aware parsing framework specifically designed to process

heterogeneous resume structures while maintaining spatial metadata.

Keyword matching fails when applicants use alternative terminology for required skills.

Transformer embeddings resolve this by mapping text into a shared semantic space, enabling context-aware comparisons regardless of phrasing.

To complement semantic vectorization, weighted set metrics evaluate skill overlap.

Unlike standard set intersection, Weighted Jaccard Similarity assigns mathematical weights to individual terms based on importance. Zhang et al. (2026) demonstrated its effectiveness in aligning text across sets with non-uniform element weights.

This study combines Transformer embeddings and Weighted Jaccard Similarity into a hybrid matching engine. Transformer vectors measure contextual alignment, while Weighted Jaccard Similarity calculates weighted skill overlap. The web platform will process resume uploads, extract candidate data, execute hybrid matching against job requirements, and display similarity scores for Systems Plus College Foundation, Inc.

1. Application Abandonment: Recruitment friction, poor communication, and complex screening trigger high candidate drop-out rates (Bullhorn, 2022).
2. Layout Diversity: Multi-column layouts and visual structures disrupt reading order reconstruction and section extraction (Werner & Laber, 2024; Zhu et al., 2025).
3. Exact Keyword Limitations: Standard keyword matching fails to identify semantically equivalent qualifications written in alternative terminology.
4. Unweighted Skill Overlap: Conventional set-intersection metrics treat all skills equally, ignoring skill criticality (Zhang et al., 2026).

Existing studies have explored layout parsing, Transformer embeddings, and Weighted Jaccard Similarity (Werner & Laber, 2024; Zhang et al., 2026; Zhu et al., 2025). However, these approaches are often studied separately indicating a methodological gap where such methods could be utilized in one study. Although visual OCR can parse layout structure and Transformer models can capture semantic meaning, limited research combines these approaches into a single web-based system that parses heterogeneous resumes, evaluates skill criticality, and calculates candidate compatibility at Systems Plus College Foundation, Inc.

1.3 Objectives of the Study

- **General Objective:** Develop a Web-based Resume-Job Matching System Using Transformer Embeddings and Weighted Jaccard Similarity at Systems Plus College Foundation, Inc. to extract structured candidate data and evaluate applicant-job alignment.
- **Specific Objectives:**
 1. Construct a visual OCR and parsing pipeline to extract structured text from heterogeneous resume layouts.
 2. Develop a Transformer embedding model to project candidate qualifications and job specifications into a unified semantic space.
 3. Build a hybrid matching module combining Transformer semantic similarity with Weighted Jaccard Similarity to compute match scores.
 4. Deploy a web platform allowing authorized users to upload resumes, input job descriptions, review parsed data, and view match analytics.
 5. Evaluate system performance regarding parsing accuracy, matching precision, and processing speed against keyword-based approaches.

1.4 Scope and Limitation of the Study

- **Scope:** This study focuses on developing a web-based candidate matching system for Systems Plus College Foundation, Inc. incorporating visual OCR, document parsing, Transformer embeddings, and Weighted Jaccard Similarity. The platform targets core metadata: contact details, education, technical skills, work experience, and certifications. A hybrid engine evaluates this data against job descriptions by computing semantic vectors and weighted skill overlap. A web dashboard presents similarity scores and matching breakdowns to support initial screening. Evaluation covers technical accuracy, parsing precision, skill-overlap metrics, and processing efficiency. The system acts strictly as a decision-support tool.
- **Limitation:** Parsing accuracy depends on document quality and structure; performance declines on corrupt, poorly scanned, handwritten, or password-protected files. No parser guarantees 100% extraction accuracy across all formats. Transformer performance remains bounded by hardware capacity and training data limits. The system does not claim to eliminate all recruitment bias. The software supports initial candidate screening only. It does not automate final hiring, conduct interviews, perform background checks, or negotiate compensation. Findings reflect conditions at Systems Plus College Foundation, Inc. and may not generalize across all institutional environments.

1.5 Significance of the Study

This study aims to improve recruitment efficiency and candidate evaluation through a web-based resume-job matching system that incorporates visual OCR, Transformer embeddings, and Weighted Jaccard Similarity. The study may benefit the following:

Human Resource Personnel and Recruiters. The study may provide an automated screening framework that reduces manual resume processing, identifies semantic qualification matches, and categorizes applicants efficiently.

Job Applicants. The study may contribute to fairer candidate evaluation by recognizing skill synonyms and contextual qualifications beyond rigid exact-keyword matching.

Systems Plus College Foundation, Inc. The study may serve as a localized institutional recruitment tool that demonstrates practical integration of Natural Language Processing, OCR, and AI-assisted screening in administrative workflows.

Future Researchers and Students. The study may serve as a reference for future research involving visual OCR parsing, Transformer-based contextual vectorization, Weighted Jaccard skill scoring, and practical full-stack AI development.

1.6 Definition of Terms

Applicant Tracking System (ATS) - A system designed to process and manage recruitment data (Merriam Webster, n.d.). In this study, it serves as the traditional keyword-matching baseline platform used to benchmark the proposed hybrid engine.

Candidate Profile - A brief statement giving key information about a person (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 993). In this study, it represents the structured applicant data, including skills, experience, and education, parsed from a resume for algorithmic evaluation.

Job Description - An account of the properties, characteristics, or requirements of something (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 336). In this study, it specifies the position duties and required skills used as the target query for candidate evaluation.

Natural Language Processing (NLP) - The capability of an artificial intelligence system to understand and process natural human language (Merriam-Webster, n.d.). In this study, it encompasses the computational text-processing techniques used to clean resume text and construct vector spaces.

Optical Character Recognition (OCR) - The process of converting printed or visual text into machine-readable digital data (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 872). In this study, it denotes the visual engine used to capture text from scanned, image-based, or non-standard PDF resumes.

Resume Parsing - To divide text into components and analyze their structural relations (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 914). In this study, it applies to the automated extraction and mapping of unstructured resume layout data into categorized database fields.

Transformer Embeddings - The mathematical mapping of one structure into another space (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 405). In this study, these represent high-dimensional numerical vectors generated by language models to capture candidate qualifications contextually.

Semantic Similarity - Relating to meaning in language and the state of being comparable (Merriam-Webster's Collegiate Dictionary, 11th ed., pp. 1131, 1170). In this study, it calculates the conceptual alignment between resume content and job specifications, regardless of exact wording.

Weighted Jaccard Similarity - A set-based similarity metric that incorporates element weights into set intersections and unions. In this study, it evaluates candidate skill overlap based on term criticality rather than unweighted keyword counts.

Resume-Job Matching - The act of bringing two entities into harmonious relation or direct correspondence (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 765). In this study, it describes the

algorithmic alignment of candidate profiles against job requirements using combined similarity metrics.

Matching Score - A numerical record or quantitative evaluation (Merriam-Webster's Collegiate Dictionary, 11th ed., p. 1108). In this study, it indicates the composite percentage generated by the system to rank candidate suitability on the web dashboard.

Visual OCR Parsing - Layout-aware processing combining visual recognition with structural document analysis. In this study, it refers to the specialized extraction pipeline engineered for multi-column, graphical, and non-standard resume layouts.

SECTION II — LITERATURE REVIEW

2.1 Introduction to the Review

This chapter reviews literature and studies related to intelligent resume screening, job recommendation, resume information extraction, semantic matching, Transformer-based embeddings, document layout understanding, and fairness in AI-assisted recruitment. The reviewed studies cover six main topics: AI-assisted recruitment and automated resume screening, job recommendation and matching algorithms, Transformer-based semantic text embeddings, document layout understanding and optical character recognition (OCR) for non-standard resumes, similarity scoring methods with term weighting, and AI recruitment implementation within the Philippine local context.

Together, these studies highlight a key gap in current research: while many tools focus on individual parts of document processing and recruitment analysis, few combine layout-aware resume information extraction, Transformer embeddings, and Weighted Jaccard Similarity into a single web-based matching workflow tailored for Systems Plus College Foundation, Inc. (SPCF). Looking at these different areas helps show why an integrated system is needed to support fair, transparent, and efficient recruitment workflows.

2.2 Review of Related Literature and Studies

2.2.1 AI-Assisted Recruitment and Resume Screening

Artificial intelligence is increasingly used across recruitment activities, including resume screening, candidate ranking, job recommendation, and decision support. A systematic review by Dadaboyev et al. (2025) based on 49 peer-reviewed articles reports that AI applications automate resume parsing and initial screening while raising concerns regarding transparency and algorithmic bias. A broader review by González-Morales et al. (2026) similarly describes machine learning and natural language processing (NLP) as increasingly important in candidate pre-selection, resume classification, job suitability prediction, and recommendation. Abhishek et al. (2025) specifically developed an AI-driven resume analysis tool, describing manual screening as time-consuming and vulnerable to oversight and inconsistent evaluation. Furthermore, ethical evaluations by Hunkenschroer and Luetge (2022) as well as Hunkenschroer and Kriebitz (2023) emphasize that an automated matching score should be transparent and serve as decision-support to human recruiters rather than silently replacing human judgment.

2.2.2 Job Recommendation and Resume-Job Matching

Job recommender systems are designed to reduce the difficulty of finding suitable jobs or candidates within large amounts of information. Çelik Ertuğrul and Bitirim (2025) reviewed 57 research articles published from 2010 to 2023, identifying content-based, collaborative, hybrid, knowledge-based, context-aware, and reciprocal approaches using user profiles, job profiles, and similarity measures. Earlier work by Roy et al. (2020) demonstrated a machine learning framework that classified resumes and ranked candidates according to a job

description using content-based recommendation, cosine similarity, and k-nearest-neighbor techniques. Chou and Yu (2020) investigated AI-based resume analysis and job recommendation, showing the effectiveness of text mining for connecting applicant information with job opportunities. Additionally, Alsaif et al. (2022) proposed a bi-directional recommendation system enabling recruiters to find suitable resumes while job seekers find suitable jobs using NLP and similarity calculations.

2.2.3 Transformer Embeddings and Semantic Matching

A major development in NLP is the use of Transformer-based language models that capture context and meaning beyond exact keyword matches. Devlin et al. (2019) introduced BERT, a bidirectional Transformer model that learns contextual representations from text. Reimers and Gurevych (2019) extended this direction with Sentence-BERT (SBERT), which produces semantically meaningful sentence embeddings that can be compared efficiently using cosine similarity. Rosenberger et al. (2025) developed CareerBERT for matching resumes to ESCO jobs in a shared embedding space, illustrating the value of semantic representations in job recommendation. Jagadeesh et al. (2025) likewise explored AI-driven resume analysis using semantic modeling and large language model feedback loops, emphasizing the importance of understanding contextual meaning over exact word matching.

2.2.4 Document Layout, OCR, and Non-Standard Resumes

Extracting information from non-standard resumes containing multiple columns, tables, graphics, and visually complex layouts requires document layout understanding. Xu et al. (2020) introduced LayoutLM, which jointly models text and layout information for visually rich documents. Appalaraju et al. (2021) developed DocFormer, a multimodal Transformer that combines text, visual, and spatial features for document understanding. Huang et al. (2022) later proposed LayoutLMv3, which uses unified text and image masking to learn cross-modal relationships for Document AI tasks. These studies indicate that document parsing and semantic matching should be treated as connected components, as incomplete extraction negatively impacts semantic matching accuracy.

2.2.5 Similarity Scoring and Weighted Jaccard

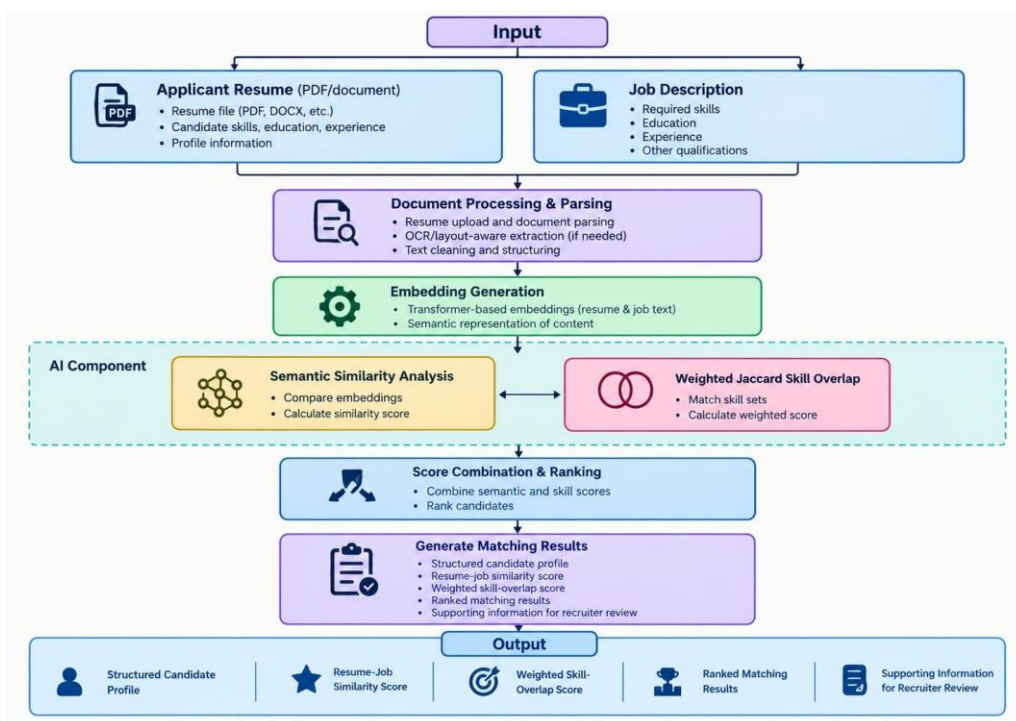
Similarity measures provide the numerical basis for comparing candidate information with job requirements. While traditional content-based recommendation systems rely on measures such as cosine similarity (Roy et al., 2020), Weighted Jaccard Similarity allows specific terms, such as required technical skills, to carry higher weight than secondary qualifications. By comparing the weighted intersection with the weighted union of terms, Weighted Jaccard Similarity provides an interpretable measure of term-level skill overlap that complements the semantic representations of Transformer embeddings.

2.2.6 Philippine and Local Recruitment Context

Philippine studies indicate that AI adoption in recruitment is an emerging area of institutional practice. Zhang (2024) examined HR professionals' perceptions of AI adoption in recruitment across two Philippine companies, highlighting that organizational success depends on how personnel perceive and use the technology. A study by Vedija and Martir (2026) among higher education institutions in Antique documented AI usage in qualification matching and resume screening while identifying data privacy, cost, and technical constraints as operational challenges. Agngarayngay et al. (2026) reported that usability, trust, and perceived fairness are crucial predictors of AI acceptability among HR professionals in Metro Manila. When applied to Systems Plus College Foundation, Inc. (SPCF), these findings demonstrate the need for a manageable, web-based tool adapted to local institutional workflows.

2.3 Conceptual Framework

The conceptual framework follows an Input-Process-Output (IPO) model, detailing the architecture and sequential data flow of the proposed recruitment screening system.



The input stage consists of two primary data sources: the applicant resume (containing the document file, candidate skills, education, experience, and profile details) and the job description (containing required skills, education, experience, and other qualifications).

The system processes this input through sequential data handling and machine learning pipeline stages. Initial document processing and parsing performs file uploads, OCR or layout-aware extraction where necessary, and text cleaning to convert unstructured text into structured formats. Next, embedding generation transforms both the resume and job text into dense semantic vector representations using Transformer-based models.

Within the core AI component, semantic similarity analysis and weighted Jaccard skill overlap evaluation operate in parallel, interacting to analyze high-level context alongside explicit, weighted skill set matching. The resulting similarity and skill overlap scores are then synthesized during score combination and ranking to produce a final candidate order. Finally, candidate details and metrics are structured into detailed matching results.

The final output provides recruiters with five structured deliverables: a structured candidate profile, a resume-job similarity score, a weighted skill-overlap score, ranked matching results, and supporting information for recruiter review.

2.4 Synthesis

The reviewed literature confirms that AI and natural language processing significantly improve recruitment workflows, including resume screening, candidate ranking, and job recommendation. Studies on text representations show that traditional keyword matching fails when skills are phrased differently, whereas Transformer models such as BERT, SBERT, and CareerBERT capture contextual meaning. Document AI models such as LayoutLM, DocFormer, and LayoutLMv3 further demonstrate that visual structure and layout must be processed alongside text content when parsing multi-column or non-standard resumes.

Research on similarity scoring demonstrates that combining semantic vector representations with Weighted

Jaccard Similarity creates a balanced framework where contextual meaning and weighted term-level skill overlap are evaluated simultaneously. Furthermore, ethical and usability studies emphasize that recruitment AI should produce transparent, interpretable outputs to build recruiter trust and prevent automated bias.

Philippine studies support the growing institutional adoption of AI in human resource management while identifying persistent challenges involving technical usability, data privacy, and trust. However, the reviewed Philippine literature does not evaluate an integrated system combining layout-aware parsing, Transformer embeddings, and Weighted Jaccard scoring specifically implemented for Systems Plus College Foundation, Inc.

Overall, no reviewed study combines layout-aware document extraction, Transformer-based semantic embeddings, and Weighted Jaccard skill-overlap scoring into a single web-based matching workflow for SPCF. The proposed study addresses this gap by developing a lightweight, transparent, web-based resume-job matching system that calculates combined similarity scores to assist recruiters during early-stage candidate evaluation.

SECTION III — METHODOLOGY

3.1 Research Design

This study employs a Developmental Research Design with a Quantitative Approach. The developmental aspect guides the design, implementation, and integration of the Web-Based Resume-Job Matching System Using Transformer Embeddings and Weighted Jaccard Similarity at Systems Plus College Foundation, Inc. The proposed system integrates layout-aware visual Optical Character Recognition (OCR), structured resume information extraction, Transformer-based semantic embeddings, and Weighted Jaccard Similarity. The descriptive-quantitative aspect evaluates the system's performance by measuring layout-parsing accuracy, semantic extraction precision, matching performance, and processing efficiency. The system's output will be benchmarked against conventional keyword-based matching approaches to determine differences in candidate processing and matching alignment.

3.2 System Development Methodology

The researchers will use the V-Model to develop and test the proposed system. The V-Model connects each development stage with a matching testing stage. The left side focuses on planning and design, while the right side focuses on testing the finished parts of the system.

The V-Model is suitable for this study because the system has several parts that must work together. These include document ingestion, layout-aware OCR, information extraction, Transformer-based semantic embeddings, Weighted Jaccard similarity computation, background processing queues, database operations, and the web interface. Each part will be planned, developed, and tested based on the requirements of the study. The system will follow these stages:

1. Requirements Analysis

The researchers will identify what the system needs to do. The system must accept unstructured resumes, extract layout-aware text via OCR, parse structured candidate information, generate semantic vector embeddings, compute similarity scores, and present explainable matching results on a web interface.

The researchers will also identify the needed hardware, software, school-layout data, crowd settings, and evaluation measures. These requirements will serve as the basis for user acceptance testing.

2. System Design

The researchers will design how the complete system will work. This stage will show how the web front-end, background worker pool, OCR engine, NLP entity parser, Transformer embedding model, similarity computation

engine, and database layers will connect.

The researchers will also identify the system inputs and outputs. Inputs may include uploaded resume files, job descriptions, and user configuration settings. Outputs may include extracted candidate profiles, layout-parsing accuracy, semantic precision, runtime processing speed, and match compatibility percentages. The complete system design will serve as the basis for system testing.

3. Architectural Design

The system will be divided into its main parts. These will include the following:

1. Web interface and user dashboard
2. Document ingestion and file validation service
3. Layout-aware visual OCR engine
4. Information extraction and entity parsing module
5. Transformer-based semantic embedding generator
6. Weighted Jaccard and vector similarity computation engine
7. Asynchronous background queue and worker pool
8. Database and profile repository

The researchers will identify how these parts will share information. For example, the ingestion service passes uploaded files to background queues, where OCR and NLP modules extract structured entities. The embedding model generates vector representations, which are passed to the similarity engine to calculate match scores. The connections between these parts will be checked during integration testing.

4. Module Design

The researchers will prepare the detailed design of each system part. The OCR module will handle multi-column, graphic-heavy, and non-standard PDF formats.

The NLP entity parser will define target entities such as skills, work history, education, and certifications. The embedding module will define vector dimensions, and the similarity engine will specify the scoring logic for Weighted Jaccard and vector distance calculations. Each module will be checked during unit testing.

5. System Development

The researchers will build the system based on the approved designs. The web interface and database operations will be created alongside background queue services.

The layout-aware OCR, entity extraction pipelines, Transformer models, and similarity engines will be integrated into the processing workflow. After the parts are developed, they will be combined into one working web-based resume-job matching system.

6. Unit Testing

Each part of the system will be tested separately. The researchers will test document upload validation, text extraction accuracy, entity parsing precision, vector embedding generation, similarity score calculations, background queue execution, and interface responsiveness.

The extracted entities and similarity scores will be verified against expected ground truth data. Any part that does not produce the expected result will be corrected and tested again.

7. Integration Testing

Integration testing will check whether the different parts work correctly when connected. The researchers will check whether the front-end successfully routes files to background queues, whether OCR text flows correctly to NLP and embedding models, and whether generated scores accurately populate the database and user dashboard.

Problems involving data transfer, timing, or connections between the system parts will be corrected before system testing.

8. System Testing

The researchers will test the complete system based on the requirements and system design. Controlled simulation runs will benchmark the proposed Transformer-Jaccard system against conventional keyword-based matching approaches. Both methods will be tested using identical sets of resumes and job descriptions.

The researchers will compare layout-parsing accuracy, semantic extraction precision, runtime processing efficiency, and false rejection rates produced by each approach. They will also test non-standard resume layouts to evaluate whether the visual OCR and semantic pipelines reduce parsing failures.

9. User Acceptance Testing

Recruitment personnel, system operators, and stakeholders will evaluate the completed system. They will answer a structured Likert-scale questionnaire based on selected ISO/IEC 25010 software-quality characteristics.

The questionnaire will measure the system's functional suitability, usability, reliability, and performance efficiency. The study will use purposive sampling because the participants will be selected based on their connection to the intended users of the system.

Participation will be voluntary, and participants may stop at any time without penalty. The questionnaire results and system performance data will be analyzed separately. The questionnaire will show how users rate the system's usability and transparency, while empirical benchmark data will show the technical performance of the OCR, embedding, and similarity modules.

In the V-Model, every development stage has a matching testing stage. User acceptance testing checks the requirements. System testing checks the complete system design. Integration testing checks whether the main parts work together. Unit testing checks each individual module. This process will help the researchers confirm that the system works as planned and meets the objectives of the study.

3.3 Requirement Specification and Analysis

This section identifies the functions and requirements of the proposed Web-Based Resume-Job Matching System. The requirements were based on the objectives of the study and the expected operation of the system. The proposed system will accept unstructured resume files, process complex layouts via visual OCR, extract structured candidate attributes, generate Transformer-based vector embeddings, compute Weighted Jaccard and semantic similarity scores, and present explainable matching results.

3.3.1 Functional Requirements

Functional requirements describe the actions that the system must perform.

1. Document Ingestion

The system must allow users to upload unstructured resume files in PDF and image formats and perform file format validation.

2. Layout-Aware OCR Extraction

The system must utilize visual OCR to extract text and maintain spatial layout context from multi-column, graphic-heavy, and non-standard resume formats.

3. Entity Parsing & Structuring

The system must parse extracted text into structured candidate entities, including technical skills, work experience, education, and certifications.

4. Semantic Vector Embedding Generation

The system must transform parsed candidate profiles and posted job descriptions into high-dimensional vector embeddings using a Transformer model.

5. Dual Similarity Computation

The system must compute semantic similarity via vector distance search and evaluate attribute/token overlap using Weighted Jaccard Similarity.

6. Calibrated Score Generation

The system must integrate vector similarity and Weighted Jaccard metrics to generate a calibrated match percentage.

7. Asynchronous Task Processing

The system must offload heavy extraction and embedding tasks to background worker queues to preserve interface responsiveness.

8. Match Result Presentation

The system must present match percentages, candidate attribute breakdowns, and comparative candidate rankings on a clear web dashboard.

9. System Performance Benchmarking

The system must record layout-parsing accuracy, semantic precision, and execution speed to enable comparison against conventional keyword-based matching baselines.

10. User Evaluation

The completed system must be available for evaluation by selected recruitment personnel and system operators through a structured questionnaire.

3.3.2 Nonfunctional Requirements

Nonfunctional requirements describe how the system should perform.

1. Usability

The controls, labels, instructions, and settings must be easy to understand. Users must be able to start, stop, and exit the simulation without difficulty.

2. Performance Efficiency

The system must process crowd movement, congestion prediction, and route updates without delays that interrupt the simulation.

3. Reliability

The system must record the required data correctly. It must also produce consistent results when the same settings and conditions are used.

4. Accuracy

The GRU predictions must be compared with the actual congestion recorded during the simulation. Routes created by A* and D* Lite must lead agents toward valid exits.

5. Compatibility

The system must work with the computer and VR equipment selected for the study. Unity, the GRU model, and the routing components must be able to exchange the required data.

6. Safety

The VR simulation must avoid unnecessary graphic violence. Users must be able to stop the simulation immediately if they feel afraid, uncomfortable, dizzy, or unwell.

7. Data Protection

Participant responses and simulation records must be stored securely. The presented results must not reveal

the identity of any participant.

8. Maintainability

The system must allow the researchers to update the school layout, crowd settings, GRU model, and routing functions when corrections are needed.

3.3.3 Level 0 Data Flow Diagram

The Level 0 Data Flow Diagram presents the proposed system as one main process. It shows the basic exchange of information between the users (Applicants/Recruiters) and the Web-Based Resume-Job Matching System.

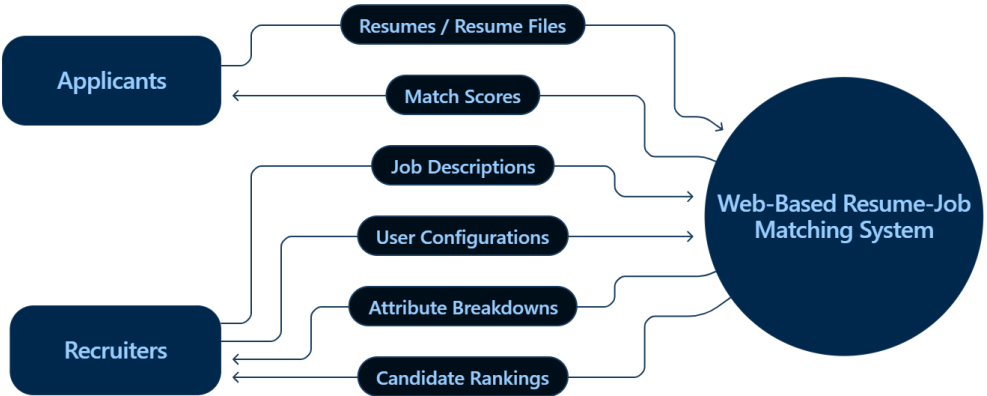
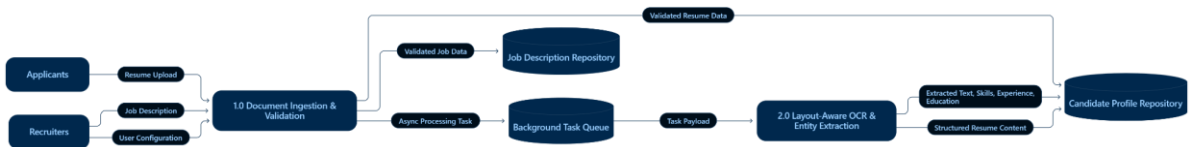


Figure 3.1. Level 0 Data Flow Diagram of the Proposed System

Figure 3.1 shows the main flow of data between the users and the proposed system. Users submit job descriptions and uploaded resumes. The system processes these inputs and produces structured candidate profiles, vector embeddings, similarity match percentages, and comparative candidate rankings.

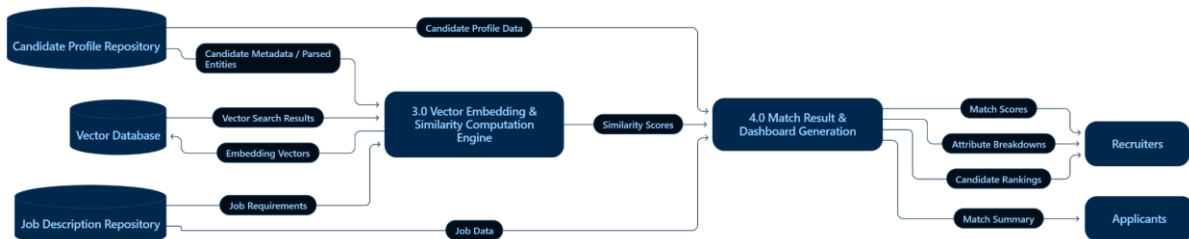
3.3.4A Level 1 Data Flow Diagram: Ingestion and Extraction

The Level 1 Data Flow Diagram divides the proposed system into its main processes. It provides a more detailed view of how the system receives, processes, stores, and presents information.



This part covers the front half of the system: upload, validation, OCR, and entity extraction.

3.3.4B Level 1 Data Flow Diagram: Embedding, Similarity, and Results



This part covers the back half: vector processing, similarity scoring, and dashboard output.

Figure 3.2. Level 1 Data Flow Diagram of the Proposed System

Figure 3.2 shows the main processes of the proposed system. The ingestion process receives uploaded files and routes them to visual OCR parsing. The entity extraction module structures candidate attributes into the candidate repository. Transformer models generate embeddings stored in the vector store, while the similarity engine computes Weighted Jaccard and vector distance scores against active job descriptions. Result summaries are then generated and displayed on the web dashboard.

3.3.5 System Flowchart

The system flowchart shows the order of activities performed by the proposed simulator. It covers the process from document upload to recording and presenting match results.

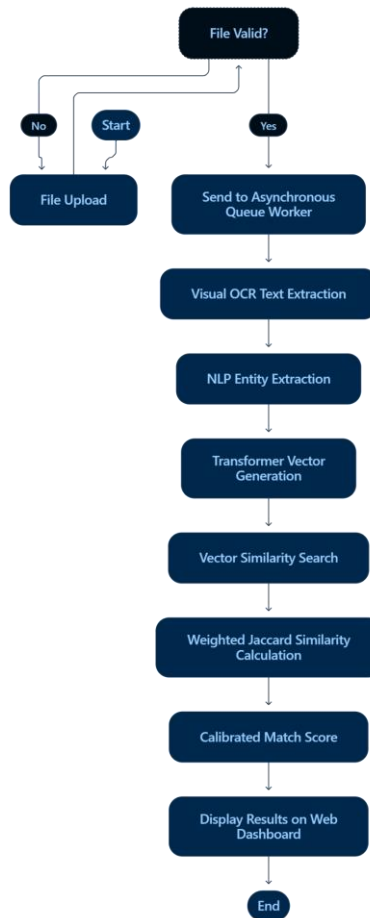


Figure 3.3. System Flowchart of the Proposed System

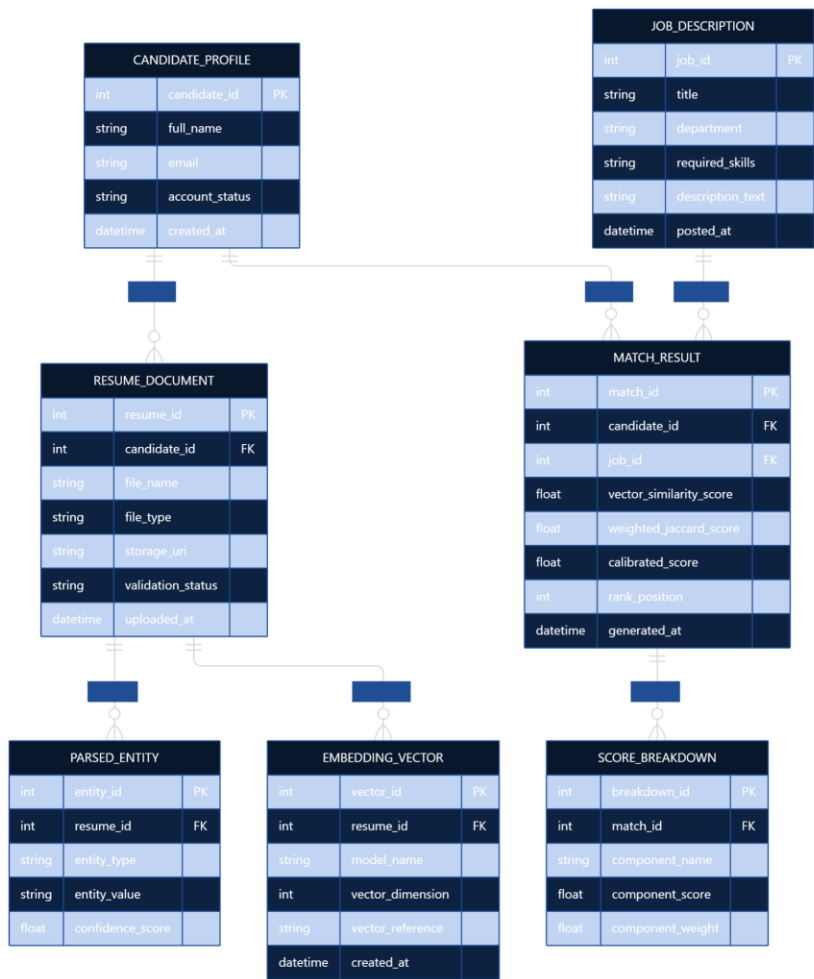
Figure 3.3 shows the main execution flow. Users upload a resume, and format validation directs the document to an asynchronous worker queue. Layout-aware OCR extracts spatial text, followed by NLP entity extraction. The Transformer model generates high-dimensional embeddings, after which vector similarity and Weighted Jaccard scores are calculated. The system integrates these metrics into a calibrated match score displayed on the candidate dashboard.

3.3.6 Entity-Relationship Diagram

The Entity-Relationship Diagram shows how the main data groups of the proposed system are connected. It provides a guide for organizing the information recorded during each processing run.

Figure 3.4. Entity-Relationship Diagram of the Proposed System

Figure 3.4 shows the relationships among Candidate Profiles, Uploaded Documents, Parsed Entities, Job Descriptions, Embedding Vectors, and Match Results. An uploaded document produces multiple parsed entities and vector embeddings. These entities and vectors are evaluated against active job descriptions to generate match result records containing compatibility scores and score breakdowns.



3.4 System Design Specification

- Logical Specification:** The system will provide a user interface for applicants and recruiters to upload resumes, post job descriptions, and view candidate rankings. Heavy computational processes, including layout parsing, visual OCR, and Transformer vector generation, are offloaded to asynchronous background queues to maintain front-end responsiveness. The module logic allows background workers to extract structured entities (skills, work history, education), generate high-dimensional vector embeddings, perform vector distance search against active job postings, and execute Weighted Jaccard Similarity calculations to output a calibrated compatibility score.
- Physical Specification:** The system will use a modern full-stack web framework integrated with a relational database for user and job metadata, a vector database for embedding storage, and an asynchronous queue service for background job processing. Transformer models and visual OCR engines will execute on a

hardware-accelerated server environment (GPU-supported) connected to client machines via standard web protocols.

3.5 Implementation

The system will be implemented following the designed modules and system flow. Development will encompass constructing the web dashboard, configuring relational and vector database schemas, setting up asynchronous processing queues, and integrating visual OCR and Transformer pipelines. The similarity engine will combine vector distance calculations with Weighted Jaccard scoring to produce transparent, explainable candidate evaluations.

3.6 Testing and Evaluation

The system will be evaluated through Unit Testing, Integration Testing, System Testing, and User Acceptance Testing (UAT). Unit testing will verify individual modules (OCR extraction, entity parsing, embedding generation), while integration testing will check data flow across file ingestion, background queues, databases, and front-end displays. System testing will compare the proposed Transformer-Jaccard approach against conventional keyword-based baselines using controlled resume datasets. Performance will be measured using layout-parsing accuracy, semantic extraction precision, runtime processing efficiency, and false rejection rates. UAT will involve recruitment personnel and system operators using a 5-point Likert-scale questionnaire based on selected ISO/IEC 25010 quality characteristics, focusing on functional suitability, usability, reliability, and performance efficiency.

REFERENCES

- Abhishek, K. L., Niranjanamurthy, M., Aric, S., Ansarullah, S. I., Sinha, A., Tejani, G., & Shah, M. A. (2025). Developing an intelligent resume screening tool with AI-driven analysis and recommendation features. *Applied AI Letters*, 6(2), e116. <https://doi.org/10.1002/ail2.116>
- Agngarayngay, C. D., Latifi, S. I., Marco, F. J. P., Tan, C. E., & Villamor-Evangelista, M. G. (2026). AI integration in screening and selection: The impact of perception of fairness, trust, and usability in the acceptability of HR employees in Metro Manila. *International Journal of Business and Social Science Research*, 7(3). <https://doi.org/10.47742/ijbssr.v7n3p2>
- Alsaif, S. A., Hidri, M. S., Ferjani, I., Eleraky, H. A., & Hidri, A. (2022). NLP-based bi-directional recommendation system: Towards recommending jobs to job seekers and resumes to recruiters. *Big Data and Cognitive Computing*, 6(4), Article 147. <https://doi.org/10.3390/bdcc6040147>
- Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 993–1003). IEEE. https://openaccess.thecvf.com/content/ICCV2021/html/Appalaraju_DocFormer_End-to-End_Transformer_for_Document_Understanding_ICCV_2021_paper.html
- Bullhorn. (2022, September 28). Poor hiring processes cause 75% of Gen Z to abandon promising job applications. <https://www.bullhorn.com/news-and-press/press-releases/gen-z-talent-survey/>
- Çelik Ertuğrul, D., & Bitirim, S. (2025). Job recommender systems: A systematic literature review, applications, open issues, and challenges. *Journal of Big Data*, 12, Article 140. <https://doi.org/10.1186/s40537-025-01173-y>
- Chanda, P., Sri, T. B., Tejasree, C., Vardhan, N. M., & Rohith, M. (2026). Resume screening system with semantic search, bias-mitigation, and explainability. *International Journal of Scientific Development and Research*, 11(2).

<https://ijsdr.org/papers/IJSDR2602024.pdf>

Dadaboyev, S. M. U., Abdullayeva, J., Abbosova, N., & Suleymenova, A. (2025). Role of artificial intelligence in employee recruitment: Systematic review and future research directions. *Discover Global Society*, 3, Article 99. <https://doi.org/10.1007/s44282-025-00246-w>

González-Morales, A., Martín-López, M. M., & Góngora-Larequi, F. (2026). A systematic review and descriptive analysis of artificial intelligence applied to recruitment and personnel selection in the present and possible future. *Discover Artificial Intelligence*, 6, Article 762. <https://doi.org/10.1007/s44163-026-01488-6>

Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4083–4091). Association for Computing Machinery. <https://doi.org/10.1145/3503161.3548112>

Hunkenschroer, A. L., & Kriebitz, A. (2023). Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. *AI and Ethics*, 3(1), 199–213. <https://doi.org/10.1007/s43681-022-00166-4>

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178(4), 977–1007. <https://doi.org/10.1007/s10551-022-05049-6>

Jagadeesh, A., Ravi Shankar, C., Narayanaswamy Patel, S., Levantesi, M., Semeraro, G., & De Luca, E. W. (2025). AI-driven resume analysis and enhancement using semantic modeling and large language feedback loops. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics* (pp. 526–538). CEUR Workshop Proceedings. <https://aclanthology.org/2025.clicit-1.51/>

Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., & Zschech, P. (2025). CareerBERT: Matching resumes to ESCO jobs in a shared embedding space for generic job recommendations. *Expert Systems with Applications*, 275, Article 127043. <https://doi.org/10.1016/j.eswa.2025.127043>

United Nations Department of Economic and Social Affairs. (n.d.-a). Goal 8: Decent work and economic growth. United Nations. <https://sdgs.un.org/goals/goal8>

United Nations Department of Economic and Social Affairs. (n.d.-b). Goal 9: Industry, innovation and infrastructure. United Nations. <https://sdgs.un.org/goals/goal9>

Vedija, M. L., & Martir, E. M. (2026). Utilization of artificial intelligence in human resource management recruitment: Its opportunities and challenges. *Iconic Research and Engineering Journals*, 9(11), 4042–4062. <https://doi.org/10.64388/IREV9I11-1718388>

Werner, M., & Laber, E. (2024). Extracting section structure from resumes in Brazilian Portuguese. *Expert Systems with Applications*, 242, Article 122495. <https://doi.org/10.1016/j.eswa.2023.122495>

Yan, Y. (2026). Transformer-based semantic embedding model for resume-job matching in intelligent talent screening. *Artificial Intelligence and Digital Technology*, 3(1), 82–91. <https://doi.org/10.70088/mvanpe51>

Zhang, P. (2024). Application of artificial intelligence (AI) in recruitment and selection: The case of Company A and Company B. *Journal of Business and Management Studies*, 6(3). <https://doi.org/10.32996/jbms.2024.6.3.18>

Zhang, Y., Qiao, M., Peng, Z., & Deng, D. (2026). Near-duplicate text alignment under weighted Jaccard similarity. *Proceedings of the VLDB Endowment*, 19(7), 1600–1613. <https://doi.org/10.14778/3801059.3801072>

Zhu, F., Yu, J., Chen, Z., Zhou, Y., Ji, J., Yang, Z., Zhang, Y., Hu, H., & Liu, Z. (2025). Layout-aware parsing meets efficient LLMs: A unified, scalable framework for resume information extraction and evaluation (arXiv:2510.09722). arXiv. <https://arxiv.org/abs/2510.09722>