CV Classification Using Machine Learning Techniques

Roxana-Teodora PERSINARU
Bucharest University of Economic Studies
Faculty of Cybernetics, Statistics and Economic Informatics
Bucharest, Romania
persinaruroxana21@stud.ase.ro

Automatic curriculum vitae (CV) classification represents a complex challenge in the field of human resources, driven by the increasing volume of applications and the need to rapidly identify relevant candidates. Manual CV evaluation is a time-consuming and error-prone process, highlighting the necessity of automated solutions based on Machine Learning and Natural Language Processing techniques.
This paper proposes a hybrid architecture for CV classification that combines semantic representations based on multilingual embeddings, adaptive linguistic rules, and semantic similarity mechanisms for candidate–job compatibility assessment.
The proposed system enables robust classification of semi-structured CVs and automatic candidate–job matching through an explainable multicriteria evaluation mechanism. Experimental results indicate improved performance compared to a baseline approach relying exclusively on heuristic rules, demonstrating the effectiveness of integrating semantic mechanisms and hybrid classification strategies. The proposed hybrid model achieved an accuracy of 79.2% and significantly outperformed the rule-based baseline approach, which obtained an accuracy of 40.0%.
Keywords: *CV classification, NLP, semantic similarity, candidate–job matching, information extraction, embeddings*

Introduction

In the context of the digitalization of recruitment processes and the increasing volume of applications, the automatic processing of unstructured textual documents has become an important research direction in the field of Natural Language Processing. Automating CV analysis enables the reduction of evaluation time and the more efficient identification of relevant candidates in modern recruitment processes.

According to Jurafsky and Martin [1], this process involves transforming heterogeneous textual content into structured data that can subsequently be used for analysis, retrieval, and integration into information systems. In practice, this requires the identification of relevant entities, the relationships between them, and temporal information.

Curriculum vitae (CVs) represent a relevant example of non-uniformly structured documents, as they contain information regarding professional experience, education, and skills in highly variable textual forms.

This heterogeneity complicates the automatic processing and comparative evaluation of candidates, particularly in recruitment scenarios involving large volumes of applications.

As discussed in [1], methods based on lexico-syntactic patterns generally provide high precision; however, they suffer from limited coverage and require the manual definition of rules.

In contrast, supervised learning approaches can achieve high performance when labeled datasets are available, but they involve significant annotation costs and may encounter difficulties generalizing across different CV formats.

In this context, the present paper proposes an automated system for CV analysis and can-

didate–job compatibility assessment based on an integrated semantic–heuristic framework. The system combines CV section classification using semantic representations, professional experience extraction, educational information identification, and skill matching through semantic similarity techniques.

Unlike conventional approaches based exclusively on fixed rules or independent statistical models, the proposed system integrates contextual semantic classification, adaptive heuristic rules, and explainable evaluation mechanisms within a unified pipeline. The architecture is designed to increase robustness when processing CVs with inconsistent layouts and to improve generalization capabilities across documents written in diverse structural formats.

Contributions of the Paper

The main contributions of this paper are the following:

• the design of a hybrid and explainable architecture for CV classification that integrates semantic embeddings, adaptive heuristic rules, and contextual validation mechanisms within a unified processing pipeline;

• the development of an adaptive information recovery mechanism based on textual microblocks, enabling the extraction of relevant information from semi-structured CVs with inconsistent or incomplete section delimitations;

• the integration of multilingual embedding-based semantic similarity techniques for candidate–job compatibility assessment, reducing dependency on exact lexical matching;

• the proposal of a multicriteria candidate evaluation mechanism that combines professional experience, education, skills, and target-job relevance into an interpretable scoring framework;

• the implementation of confidence-based decision fusion strategies that dynamically combine Machine Learning predictions with heuristic validation rules in ambiguous classification scenarios;

• the experimental validation of the proposed hybrid approach through comparative evaluation against rule-based and TF-IDF-based baseline methods.

The proposed architecture aims to improve robustness, interpretability, and generalization capability when processing CVs with inconsistent layouts and heterogeneous textual structures.

Related Work

Modern transformer-based Natural Language Processing models, such as BERT, introduced contextual text representations obtained through bidirectional context learning within a sentence, leading to state-of-the-art performance across multiple NLP tasks [2]. A relevant extension is Sentence-BERT, which enables the generation of sentence-level embeddings optimized for efficient semantic comparison using cosine similarity, significantly reducing computational cost compared to standard BERT [3].

In the field of automated recruitment, traditional methods based on TF-IDF and classification algorithms such as Support Vector Machines (SVM) are frequently used for CV classification, achieving competitive performance on certain datasets. However, these methods are limited in their ability to capture complex semantic relationships compared to transformer-based models [4].

Li et al. [5] proposed context-aware transformer models for both CV classification and candidate–job matching tasks. Their approach enables the estimation of semantic compatibility between resumes and job descriptions, while also supporting experience-level classification in real-world recruitment scenarios.

In addition, end-to-end systems such as Smart-Hiring integrate CV information extraction (skills, experience, education) with semantic matching within a unified and explainable pipeline. These systems provide

interpretability mechanisms and are evaluated on real-world data, demonstrating the effectiveness of integrated approaches in recruitment processes [6].

At the same time, hybrid systems combine NLP techniques, machine learning methods, and semantic similarity mechanisms for candidate classification and evaluation, providing a compromise between purely rule-based approaches and fully deep learning–based methods.

However, most existing approaches treat the stages of classification, information extraction, and matching separately, or fail to achieve an optimal balance between accuracy, interpretability, and generalization.

Table 1. Comparison of the Main Methods Used for CV Classification

Note: Ref. indicates the bibliographic references corresponding to the analyzed studies.

Ref.	Method	Task	Advantage	Limitation
[2]	Bidirectional Transformer	Text classification	Rich contextual representations	High Computational cost
[3]	Siamese BERT	Semantic similarity	Efficient for Semantic similarity (faster than standard BERT)	Performance depends on task-specific fine-tuning
[7]	TF-IDF + SVM	CV classification	Simple and interpretable	Limited ability to capture semantic relationships
[5]	Context-aware Transformer	CV classification + candidate–job matching	Models CV–job relationships	High training cost and large data requirements
[6]	End-to-end NLP Pipeline	Information extraction + candidate–job matching	Explainable and comprehensive system	Error propagation between components

Table 1 highlights the advantages and limitations of the main approaches presented in the literature, justifying the need for a hybrid solution that combines their respective benefits.

Although existing approaches achieve competitive results in individual tasks such as CV classification or semantic matching, most solutions treat the stages of information extraction, classification, and semantic evaluation separately. In contrast, the proposed system integrates these components into a unified and explainable pipeline capable of handling irregularly formatted CVs while combining contextual semantic classification with adaptive heuristic rules.

Limitations of Existing Approaches

Existing approaches for CV classification present several limitations. Systems based exclusively on linguistic rules are sensitive to structural and vocabulary variations, encountering difficulties when generalizing across differently formatted CVs.

In contrast, approaches relying entirely on deep learning models require large volumes of labeled data and provide limited interpretability.

Furthermore, many systems treat section classification, information extraction, and semantic matching as separate tasks, which may lead to error propagation between stages and reduce overall robustness.

Another important limitation is represented by structurally inconsistent CVs or documents without clear section delimitations, where traditional header-based parsing methods become insufficient.

To address these issues, the present paper proposes a hybrid architecture that integrates semantic classification, heuristic rules, and adaptive information recovery mechanisms.

Methodology

The proposed system is designed as a modular architecture organized as a pipeline consisting of multiple successive stages, each responsible for extracting, structuring, and evaluating relevant information from the document.

The workflow begins with text extraction from PDF files, followed by text normalization and cleaning. The text is subsequently segmented into smaller units associated with relevant CV sections (experience, education, skills, personal information). Based on these sections, the system extracts essential information and computes compatibility scores. The final result is obtained through the aggregation of partial scores corresponding to professional experience, education, skills, and relevance to the target job.

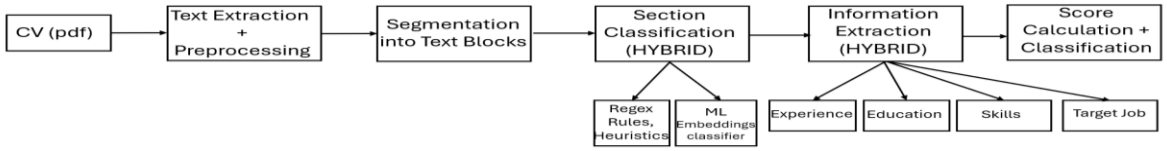

Figure 1. Architecture of the Proposed CV Classification System

Figure 1 illustrates the integration of semantic and heuristic components within the same processing pipeline. The modular organization enables the extension of the system with additional classification methods and contributes to improved adaptability when processing CVs written in non-uniform document layouts.

Unlike traditional CV parsing pipelines, the proposed architecture introduces an adaptive segmentation mechanism based on textual microblocks and contextual semantic validation, designed to reduce dependence on the explicit structure of the document.

Algorithm 1. Hybrid CV Classification Pipeline

Input:
CV document
Job requirements
Output:
Candidate class
1. Extract textual content from PDF
2. Apply normalization and text repair rules
3. Segment document into semantic text blocks
4. Detect and classify CV sections
5. Extract experience, education and skills
6. Compute semantic similarity scores
7. Aggregate weighted evaluation scores
8. Generate final candidate classification

Data Preprocessing

The preprocessing stage plays an essential role in reducing the noise introduced by the heterogeneous structure of CVs and by the text extraction process from PDF documents. CVs are written in different formats, using inconsistent delimiters, various symbols, and non-uniform information organization. In addition, the PDF extraction process may introduce further errors, such as incorrect character concatenation, delimiter loss, or inconsistent section separation.

Text is extracted from PDF files using the pdfplumber library by processing each page and concatenating the resulting textual content.

To reduce errors specific to automatically extracted documents, the system applies an additional text repair stage based on regular expressions and heuristic rules.

A relevant example is the correction of incorrect concatenations between alphabetic and numeric characters, frequently encountered in CVs converted from PDF documents:

```
text = re.sub(r'([a-zA-ZăâîșțĂÂÎȘȚ])(\d)',
r'\1 \2', text)
text = re.sub(r'(\d)([a-zA-ZăâîșțĂÂÎȘȚ])',
r'\1 \2', text)
```

This mechanism enables the correct separation of expressions such as “Python3” or “20235years”, which may occur as a result of automatic text extraction. In addition, different variations of dash delimiters (“–”, “—”, “−”) are processed and standardized into a unified format to ensure consistency during subsequent processing stages.

Table 2. Example of Textual Normalization

Raw Input	**After Normalization**
Python3Developer	Python 3 Developer
20235yearsexperience	2023 5 years experience
HR-Specialist	hr specialist

Table 2 illustrates how normalization mechanisms reduce textual inconsistencies and contribute to obtaining more stable semantic representations for subsequent processing stages.

Subsequently, the text is normalized through Unicode transformations, diacritic removal, lowercase conversion, and whitespace standardization:

```
text = unicodedata.normalize("NFKD", str(text))
text = text.encode("ascii", errors="ignore").decode("utf-8")
text = text.lower()
text = re.sub(r'[\t]+', " ", text)
```

This stage reduces lexical variations and enables the robust application of semantic classification and embedding-based similarity mechanisms using multilingual representations.

To improve semantic consistency, the system also introduces additional contextual normalization mechanisms for job positions, departments, and skills. Consequently, different variants of the same concepts (for example, “HR Specialist” and “Specialist HR”) are mapped to a common representation. This approach reduces semantic fragmentation and improves the accuracy of subsequent classification and candidate–job matching stages.

By integrating text repair mechanisms, linguistic normalization, and semantic standardization, the preprocessing stage contributes to improved reliability of the system when handling non-uniformly formatted CVs and documents extracted from different sources.

Text Segmentation

After the preprocessing stage, the document is progressively segmented into textual units that are subsequently used for semantic classification and relevant information extraction. Segmentation represents a critical stage in CV processing, as these documents do not follow a standardized structure and employ highly heterogeneous information organization patterns.

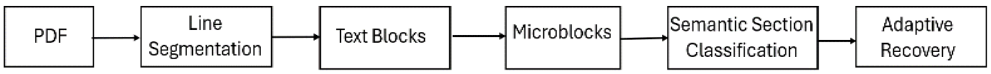

Figure 2. *Information Segmentation and Semantic Recovery Pipeline*

Figure 2 presents the overall workflow of the information segmentation and recovery mechanism used in the proposed system. The pipeline progressively transforms the PDF document into compact semantic units that are subsequently used for section classification and relevant information extraction. The introduction of the microblock-based mechanism increases robustness when processing partially structured CVs and documents without explicit section delimitations.

In the first stage, the text is divided into individual lines, while empty lines and redundant sequences are removed.

Subsequently, the lines are grouped into compact textual blocks designed to preserve

the contextual coherence of the information associated with a particular section.
To identify the logical structure of the document, the system employs a heuristic mechanism for section header detection. This mechanism analyzes features such as line length, word count, and lexical structure in order to identify expressions that may represent section titles:

```
if len(words) <= 3 and len(line) <= 35:
    if re.search(r'^\w[\w\s&/-]*$', line):
        return True
```

The detected headers are then compared against an extended set of semantic aliases for sections such as Experience, Education, Skills, and Projects. This approach enables the identification of different naming variants commonly used in real-world CVs, including formulations in both Romanian and English.
To reduce dependence on the explicit presence of section headers, the system introduces an additional mechanism based on textual microblocks. These are generated by aggregating a small number of consecutive lines, thereby preserving the local contextual information:

```
chunk = lines[i:i + window]

if chunk:
    blocks.append(normalize_text("
".join(chunk)))
```

Microblocks are used as an adaptive information recovery mechanism for poorly structured or non-uniformly formatted documents. In such situations, classification based exclusively on headers may become insufficient, as relevant information is unevenly distributed or embedded within ambiguous sections.
The system activates recovery mechanisms that analyze microblocks using heuristic rules and semantic classification in order to identify relevant sections and fragments that were not detected during the initial stage.
This mechanism is used both for CV section recovery and for the identification of professional experience and educational information.
By integrating hierarchical segmentation, adaptive header detection, and microblock-based mechanisms, the system achieves improved adaptability when processing structurally inconsistent CVs and reduces dependence on the explicit formatting of the document. Compared to traditional approaches based exclusively on fixed header identification, the proposed mechanism enables more robust processing of incomplete or non-uniformly formatted CVs.

Section Classification

The identification of relevant CV sections is performed through a hybrid architecture that combines embedding-based semantic classification, heuristic rules, and adaptive contextual validation mechanisms. This approach aims to reduce dependence on the explicit structure of the document and to improve reliability when processing non-uniformly formatted CVs.
In the initial stage, the system attempts direct section identification using an extended set of semantic aliases associated with frequently encountered CV headers. Both Romanian and English formulations are supported, including “Experience,” “Professional Experience,” “Education,” “Skills,” and “Projects.”
For textual blocks that do not contain explicit headers, the system applies semantic classification based on embeddings generated using the multilingual paraphrase-multilingual-MiniLM-L12-v2 model from the SentenceTransformers library. Textual fragments are transformed into dense vector representations capable of capturing semantic similarities between differently formulated expressions with similar meanings.

```
embeddings = embed-
```

```
ding_model.encode(blocks)
embeddings = sec-
tion_scaler.transform(embeddings)
preds = section_model.predict(embeddings)

probs = sec-
tion_model.predict_proba(embeddings)
labels = sec-
tion_labels.inverse_transform(preds)
```

Textual fragments are transformed into semantic embeddings using the multilingual SentenceTransformers model.
The resulting vectors are standardized and used as input for a Logistic Regression classifier implemented with scikit-learn, using the class_weight="balanced" parameter to reduce class imbalance effects.
In parallel, the system applies a set of heuristic rules and lexical markers specific to each category. These rules aim to identify expressions relevant to professional experience, education, technical skills, or personal information.
The heuristic approach is used both for validating semantic classifications and for recovering information in ambiguous or poorly structured documents.
The final decision is obtained through a fusion mechanism controlled by confidence thresholds. High-probability predictions are accepted directly, while low-confidence classifications are additionally validated using linguistic rules and contextual analysis:

- if $p \geq 0.72$, the semantic classification is accepted directly;
- if $0.50 \leq p < 0.72$, the final decision is obtained by combining semantic classification with heuristic rules;
- if $p < 0.50$, heuristic rules are given priority.

The decision thresholds were established empirically in order to achieve a balance between classification reliability and the reduction of false-positive errors in heterogeneous CVs.
To improve contextual interpretation, the system additionally employs a Named Entity Recognition (NER) component based on spaCy. This component functions as an auxiliary mechanism for identifying relevant entities such as persons, organizations, or professional positions, without representing the primary decision-making mechanism.

Compared to traditional approaches based exclusively on rules or fixed header identification, the proposed architecture enables more robust classification of CVs with inconsistent layouts and improves generalization capability across documents written in different formats.

Machine Learning Models

The system employs three independent supervised classifiers:

1. CV section classification;
2. professional experience entry detection;
3. educational domain estimation.

For CV section classification, a dataset consisting of 360 manually labeled textual fragments was constructed and distributed across 9 relevant semantic categories (Experience, Education, Skills, Projects, etc.). For educational domain classification, an additional dataset containing 270 labeled examples was used.
The training data were manually annotated and designed to include variations in structure, language, and information organization specific to real-world CVs.
Textual fragments were transformed into semantic embeddings using the multilingual paraphrase-multilingual-MiniLM-L12-v2 model from the SentenceTransformers library. The resulting vector representations were standardized using StandardScaler and used as input for Logistic Regression classifiers. Model evaluation was performed using an 80/20 train–test split.
The models are further integrated into a hybrid architecture that combines semantic classification with heuristic rules and adap-

tive contextual validation mechanisms in order to increase stability when processing semi-structured and non-uniformly formatted CVs.

Logistic Regression was selected due to its high interpretability and stable behavior in scenarios based on dense embeddings and standardized vector representations.

The datasets contain task-specific textual fragments:

- for section classification: textual blocks labeled as Experience, Education, Skills, Projects, etc.;
- for professional experience detection: job titles, role descriptions, and fragments associated with professional activities;
- for education classification: study programs, diplomas, and specializations labeled according to domains such as technical, economic, or administrative fields.

Relevant Information Extraction

Professional Experience

Professional experience identification represents one of the most important components of the system, as information associated with professional activities is frequently written in diverse CV structures and without standardized delimitations.

For each candidate textual fragment, the system initially applies a semantic classification model that estimates the probability that the fragment represents a professional experience entry. Candidate blocks are selected primarily from the Experience section. To reduce dependence on the explicit structure of the CV, in situations where this section is absent or insufficiently populated, the system activates adaptive recovery mechanisms and extends the search to the Summary, Projects, and Other sections.

The model predictions are then validated through contextual mechanisms designed to identify indicators specific to professional experience, such as occupational roles, temporal intervals, or explicit durations.

This approach reduces false-positive classifications and increases system stability when processing non-uniformly formatted CVs.

Total professional experience is estimated by combining explicit durations and temporal intervals automatically extracted from the document. To avoid overestimation of professional experience, the system applies an overlapping interval merging mechanism.

Through this strategy, overlaps between identified professional periods are eliminated, resulting in a more realistic estimation of total experience.

Compared to approaches based exclusively on fixed section identification, the proposed mechanism enables more robust extraction of professional experience from partially structured CVs and incomplete documents.

Education

Educational information is extracted primarily from the Education section, while in its absence, the information is recovered using semantic rules and linguistic markers specific to the academic domain.

To estimate the field of study, the system employs a semantic classification model based on multilingual embeddings. This approach enables the identification of differently formulated specializations with similar semantic meanings.

Education levels (Bachelor's, Master's, PhD) and educational institutions are identified using regular expressions and dedicated lexical rules. In the case of universities, the system applies additional cleaning and deduplication stages to eliminate redundant or incomplete variants of the detected institution names.

Educational compatibility is estimated based on the probabilities generated by the semantic model for the relevant domain and is later integrated into the candidate's final score.

Skills and Semantic Matching

Skill identification employs a hybrid approach that combines exact matching, semantic expansion, and embedding-based contextual similarity.

Skills are extracted primarily from the Skills section; however, the system can also identify relevant terms in other CV sections, such as professional experience or personal projects.

For semantic similarity evaluation, the system generates vector embeddings using the multilingual SentenceTransformers model and computes the cosine similarity between the target skills and the textual fragments extracted from the CV.

The semantic similarity is computed using the cosine similarity coefficient:

$$Sim(A, B) = \frac{A \cdot B}{|A||B|}$$

Where A and B represent the embedding vectors associated with the job requirements and the CV fragments.

A skill is considered semantically identified if the similarity score exceeds the empirically determined threshold of 0.58. This approach enables the detection of relevant competencies even when they are expressed indirectly or formulated differently from the original job requirements.

By combining lexical matching with semantic similarity, the system reduces dependency on exact wording and improves generalization across differently structured CVs.

Target Job Matching

Relevance to the target job is evaluated by combining exact matching, synonym-based matching, and semantic similarity, using a threshold of 0.56. The score is adjusted through contextual bonuses when indicators of real professional experience are identified (for example, temporal intervals or organization names).

Score Calculation and Final Classification

The candidate's final score is calculated using a multicriteria mechanism based on the weighted aggregation of the main evaluated components:

$$Score = S_{exp} + S_{edu} + S_{kw} + S_{job}$$

where:

- $S_{exp} \in [0,40]$ represents the professional experience score;
- $S_{edu} \in [0,15]$ represents the education score;
- $S_{kw} \in [0,30]$ represents the skills score;
- $S_{job} \in [0,20]$ represents the target job relevance score.

The theoretical maximum score is 105 points, since target job relevance is introduced as an additional semantic validation component.

After calculating the aggregated score, the system applies adaptive adjustments: the score is penalized when relevance to the target job is low and may receive a bonus in cases of strong semantic matching.

This stage reflects the logic implemented in the code, where the total score is reduced if target_job_score < 0.35 and increased if target_job_score >= 0.70.

Final classification is performed through a multicriteria heuristic mechanism that takes into account:

- the total score;
- the degree of skill coverage;
- semantic relevance to the target job;
- the ratio between identified experience and required experience.

The final result assigns the candidate to one of the following classes:

- suitable;
- partially suitable;
- unsuitable.

By employing multicriteria aggregation, semantic similarity, and explainable rules, the system achieves a high level of interpretability and enables the justification of automatically generated decisions.

Results and Evaluation

To evaluate the proposed system, a test set consisting of 120 manually labeled CVs was constructed, evenly distributed across three classes: UNSUITABLE, PARTIALLY SUITABLE, and SUITABLE (40 examples for each class). The CVs were collected from synthetic sources and manually designed examples intended to reflect variations in structure, language, and professional experience commonly encountered in recruitment processes. The dataset was intentionally designed to maximize structural diversity and semantic variability across CV formats.

Although the use of a synthetic dataset enables controlled class distribution and the simulation of diverse recruitment scenarios, extending the evaluation to a larger number of real-world CVs could improve the system's generalization capability.

The evaluation was performed by comparing the proposed hybrid model with a baseline approach based exclusively on heuristic rules.

Evaluation Metrics

System performance was measured using the following metrics: accuracy, balanced accuracy, precision, recall, and F1-score.

Macro F1-score was selected because it treats each class equally and provides a more relevant overview of performance in multiclass scenarios. Since the test set is balanced, the values for accuracy and balanced accuracy are identical. For the same reason, Macro F1 and Weighted F1 values are very similar, as each class contributes equally to the overall evaluation.

Table 3. Overall Performance of the Evaluated Systems

Model	Accuracy	Balanced Accuracy	Macro F1	Weighted F1
Hybrid Model	0.792	0.792	0.763	0.763
Baseline (Rule-Based)	0.400	0.400	0.413	0.413

Table 3 highlights a significant performance improvement achieved by the hybrid model. Compared to the baseline approach, the proposed model obtains an increase of 39.2 percentage points in accuracy and approximately 35 percentage points in Macro F1-score. This result confirms that the integration of section classification, semantic similarity, and heuristic rules produces a more robust CV evaluation process than an approach based exclusively on rules.

The performance improvement of the hybrid model is mainly driven by the integration of contextual semantic representations and adaptive classification mechanisms. Embedding-based semantic similarity enables the identification of relationships between candidate skills and job requirements even in the presence of lexical or structural variations across CVs.

Class-Level Performance

For a more detailed analysis, Table 4 presents the precision, recall, and F1-score values for each class in the case of the hybrid model.

Table 4. Class-Level Performance of the Hybrid Model

Class	Precision	Recall	F1-score
UNSUITABLE	0.755	1.000	0.860
PARTIALLY SUITABLE	1.000	0.400	0.571
SUITABLE	0.765	0.975	0.857

Table 4 shows that the hybrid model achieves the best results for the extreme classes, UNSUITABLE and SUITABLE. For the UNSUITABLE class, the recall value of 1.000 indicates that all examples in this

category were correctly identified. Similarly, the SUITABLE class is recognized very effectively, with a recall of 0.975 and an F1-score of 0.857.

The most challenging class is PARTIALLY SUITABLE, for which the model achieves a recall of only 0.400, although precision reaches 1.000. This result indicates that the model is conservative when assigning this label: whenever it predicts the PARTIALLY SUITABLE class, the prediction is correct, but a significant number of intermediate examples are classified into the extreme categories.

This behavior is understandable, as this class contains ambiguous profiles positioned at the boundary between unsuitable and suitable candidates. The results suggest that distinguishing between partially suitable and fully suitable candidates represents a more complex semantic problem than the classification of clear extreme cases.

Confusion Matrix Analysis

For a more detailed analysis of the performance of the evaluated systems, confusion matrices were constructed for both the hybrid model and the rule-based baseline approach.

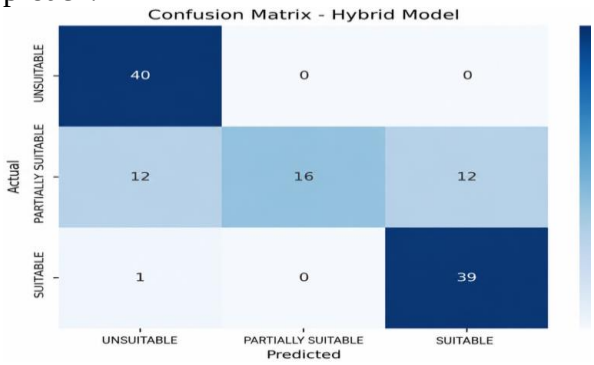

Figure 3. Confusion Matrix for the Hybrid Model

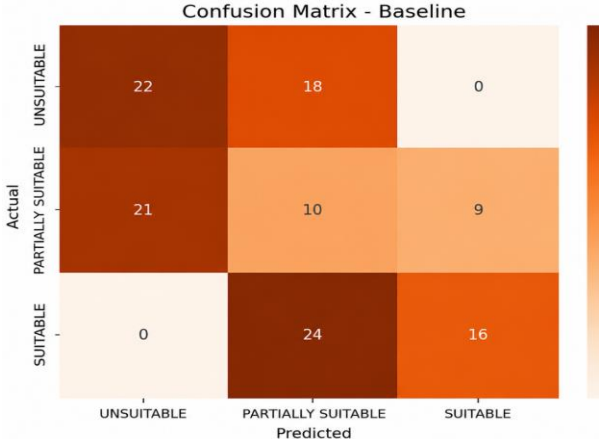

Figure 4. Confusion Matrix for the Baseline Model

Figures 3 and 4 highlight significant differences between the performance of the hybrid model and that of the rule-based baseline. The hybrid model achieves very strong results for clear cases: it perfectly classifies all UNSUITABLE examples (40 out of 40) and almost all SUITABLE examples (39 out of 40). Most difficulties occur for the PARTIALLY SUITABLE class, where only 16 out of 40 examples are correctly classified, while the remaining samples are evenly distributed between the extreme classes (12 classified as UNSUITABLE and 12 as SUITABLE).

These results indicate that the model handles obvious cases effectively but encounters difficulties in intermediate scenarios.

In contrast, the rule-based baseline demonstrates considerably weaker performance, with numerous confusions between classes and a lower number of correct classifications. For example, for the SUITABLE class, the baseline correctly classifies only 16 out of 40 examples, while for the PARTIALLY SUITABLE class only 10 out of 40 examples are correctly identified. The results confirm that simple rules and keyword-matching approaches are insufficient for capturing the complexity of information contained in CVs.

Classification Error Analysis

The most frequent classification errors occurred for CVs belonging to the PARTIALLY SUITABLE class, where the distinction between relevant and irrelevant skills is less clearly defined.

In certain cases, CVs containing general skills semantically similar to job requirements received higher scores than warranted by the actual relevance of the candidate's professional experience.

Additionally, certain very short CVs or documents written in highly non-uniform structures reduced the performance of the segmentation and section classification mechanisms. These results highlight the need for additional normalization mechanisms and for extending the dataset to cover a wider variety of scenarios.

Conclusions

The developed system demonstrates that the integration of multiple methodological approaches can lead to more efficient CV analysis, overcoming the limitations of purely rule-based systems. Experimental results indicate a strong capability to distinguish clear cases, as well as a significant improvement in overall performance.

An important aspect of the proposed solution is the way in which relevant information extracted from the document is combined, enabling a structured evaluation based on criteria such as professional experience, education, and skills.

In addition, the scoring mechanism contributes to the transparency of the final decision, facilitating the understanding of how candidates are classified.

The detailed analysis highlights that the main difficulties arise in the case of intermediate profiles, where the boundaries between classes are less clearly defined. These situations reflect the complexity of real-world recruitment processes and indicate the need for additional methods to refine decision-making in ambiguous cases.

As future work, extending the dataset and including more diverse examples, together with the exploration of more advanced models for capturing semantic relationships, may contribute to improving performance and increasing generalization capability.

The system can be integrated into human resource software applications for automating pre-screening and preliminary candidate evaluation processes, contributing to reduced analysis time and increased consistency in recruitment workflows. Furthermore, integrating the system into a broader application environment could more clearly demonstrate its usefulness in real-world recruitment scenarios.

Acknowledgment

The author expresses sincere gratitude to Professor Ion Lungu for his guidance, support, and valuable suggestions provided throughout the development of this paper.

References

[1] D. Jurafsky and J. H. Martin, "*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*", 3rd ed. Stanford University, draft version, 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/pdf/1810.04805

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP, 2019. [Online]. Available: https://arxiv.org/pdf/1908.10084

[4] N. Firdaus, B. K. Riasti, and M. A. Safi'ie, "Comparative Study of Transformer-

Based Models for Automated Resume Classification," 2025. [Online]. Available: https://www.researchgate.net/publication/398272144_COMPARATIVE_STUDY_OF_TRANSFORMER-BASED_MODELS_FOR_AUTOMATED_RESUME_CLASSIFICATION

[5] C. Li, E. Fisher, R. Thomas, S. Pittard, V. Hertzberg, and J. D. Choi, "Competence-Level Prediction and Resume & Job Description Matching Using Context-Aware Transformer Models," arXiv preprint arXiv:2011.02998, 2020. [Online]. Available: https://arxiv.org/pdf/2011.02998

[6] K. Khelkhal and D. Lanasri, "Smart-Hiring: An Explainable End-to-End Pipeline for CV Information Extraction and Job Matching," 2025. [Online]. Available: https://arxiv.org/pdf/2511.02537

[7] I. Ali, "Resume Classification System Using Natural Language Processing and Machine Learning Techniques," 2022. [Online]. Available: https://www.researchgate.net/publication/367795195_Resume_Classification_System_using_Natural_Language_Processing_and_Machine_Learning_Techniques

Roxana-Teodora Persinaru is currently pursuing a Master's degree in Databases within the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies, where she also obtained her Bachelor's degree in Economic Informatics. Her research interests include Natural Language Processing, Machine Learning, database systems, Human Resource Information Systems, and intelligent document processing.