# A Hybrid OCR-XGBoost-Transformer Pipeline for Resume Parsing with Spatial-Semantic Integration

Rachid Ed-Daoudi[1*], Fatima Zahra Zakka[2], Mouslime Ouqassou[1], Badia Ettaki[1]

E-mail: rachid.ed-daoudi@uit.ac.ma, fzakka@esi.ac.ma, mouqassou@esi.ac.ma, bettaki@esi.ac.ma.
* Corresponding author

[1]LyRICA: Laboratory of Research in Computer Science, Data Sciences and Artificial Intelligence, School of Information Sciences Rabat-Instituts, Rabat, Morocco
[2]Knowledge and data engineer, School of Information Sciences Rabat-Instituts, Rabat, Morocco

**Keywords:** resume information extraction, hybrid AI solution, optical character recognition, XGBoost, transformers

**Received:** June 5, 2025

*This study addresses the automation of resume information extraction using a hybrid Artificial Intelligence (AI) framework that integrates Optical Character Recognition (OCR), Machine Learning, and Deep Learning techniques. The system operates in three stages: text extraction using PaddleOCR, resume section classification via XGBoost, and semantic entity recognition using a Transformers-based Named Entity Recognition (NER) model. The dataset consists of 200 French resumes collected in PDF format and annotated for ten resume section classes and multiple named entities. Evaluation was conducted using standard multi-class classification metrics including accuracy, precision, recall, and F1-score. Experimental results show that XGBoost achieved 96.5% accuracy in section classification, while the Transformers model attained 82% accuracy in semantic entity extraction. This dual-stage pipeline captures both spatial and semantic structures of resumes, offering improved accuracy and adaptability over traditional parsing approaches.*

*Povzetek: Članek predstavlja hibridno rešitev OCR-XGBoost-Transformer za avtomatizirano ekstrakcijo podatkov iz življenjepisov. Sistem dosega visoko točnost pri razvrščanju razdelkov z XGBoost in pri semantičnem prepoznavanju entitet s transformerjem.*

## 1 Introduction

In an unpredictable and complex business environment, it is important that organizations aim to realize the potential offered by the recruitment phase. Organizations are in a ceaseless race to find new talent to support their teams and corporate competitiveness. The reality is that collecting candidate information from resumes is often difficult to achieve [1].

Recruiters are required to read and analyse candidate resumes manually for the information they need. This manual practice is full of disadvantages. First, it is time consuming and a labor-intensive activity for recruiters who have to read many resumes and work through a lot of information. As a result, recruiters have to deal with work overload, sometimes delaying the whole recruitment process. Therefore, an emerging technology to automate the information extraction process can be considered a rational way to control and presumably speed up a major process in recruitment [2]. The central question of this research is: How can the automation of information extraction from resumes be achieved with new technologies?

The CV parsing technology converts resume data from free-form into structured format. This conversion facilitates the storage, synthesis, and processing of information contained in resumes, thus enabling their use by software and computer systems [3]. Several parsing approaches are commonly used. Keyword-based parsers are prototypes of faster and more accurate parsers. These simplistic parsers search for specific words, key phrases, and patterns in resume text. However, this approach is prone to errors (with an accuracy rate of about 70%) as words can have multiple contexts within a resume [4].

Grammar-based parsers rely on grammatical rules to interpret information. These relatively complex parsers require manual input during the coding process. When coding is done by a skilled linguistic engineer, they can analyze a resume quite accurately. However, if manual configuration is not done correctly, grammar-based parsers can be inaccurate (with an accuracy rate of about 90%) [5].

Statistical parsers use numerical models of text to identify key elements of a resume. To be accurate, statistical parsers must be trained on a large number of resumes containing all the information to be extracted. In terms of accuracy, statistical parsers fall between keyword-based parsers and grammar-based parsers [6].

AI-based parsers use machine learning and artificial intelligence techniques. These models can improve over time by analyzing more information. AI-based parsers offer an extremely high level of accuracy compared to other CV parsing techniques available on the market [7].

Recent applications combine OCR, Computer Vision, and Natural Language Processing (NLP) techniques to advance the capabilities of resume information extraction from various formats and structures [8].

Despite advances in resume parsing technologies, existing solutions still face significant challenges in effectively handling the spatial and semantic aspects of resume documents simultaneously. Current approaches either focus on visual structure or textual content, but rarely integrate both dimensions effectively. Additionally, most commercial systems rely on rule-based methods with predefined templates, limiting their ability to process diverse resume formats and structures. There remains a need for adaptive, high-accuracy solutions that can understand both document structure and extract meaningful entities while maintaining contextual relationships across different resume sections [9].

To better position the proposed contribution, Table 1 presents a structured comparison of existing studies on resume parsing. It outlines the datasets used, methodological approaches, performance levels, and key limitations of each system. This comparative summary highlights the need for a unified system that integrates both spatial and semantic understanding of resume content.

Table 1: Summary of Related works in resume information extraction

| Ref. | Dataset Used | Method Type | Key Techniques | Accuracy / Performance | Limitations |
|---|---|---|---|---|---|
| [1] | Proprietary HR docs | Rule-based | Heuristics, Templates | Not reported | Format-dependent, low adaptability |
| [2] | Internal HR systems | Rule-based | Digital workflows, automation | Not reported | No semantic modeling, template limitations |
| [3] | 60 resumes | ML-based | Summarization, Entity extraction | ~85% accuracy | No spatial modeling, weak generalization |
| [4] | Not specified | Mixed (Keyword + ML) | NLP, keyword matching | ~70% accuracy | Poor contextual understanding |
| [5] | Literature-based | Rule-based (Survey) | Chronological parsing, analysis | N/A | No experimental validation |
| [6] | OCR-only docs | OCR | Text image recognition | ~85% OCR accuracy | No classification or entity recognition |
| [7] | Business resumes | DL-based | OCR, Deep Learning pipeline | ~90% accuracy | No spatial-semantic integration |
| [8] | English CVs | NLP + ML | NLTK-based entity recognition | Not specified | No section classification, shallow analysis |
| [9] | Polish IT resumes | Rule + ML | Section classification, heuristics | ~88% F1-score | Not end-to-end, limited semantic modeling |

As the table shows, while various parsing methods have been explored, most fail to simultaneously address spatial layout and deep semantic content. This motivates the currect hybrid OCR–XGBoost–Transformer pipeline, designed to provide accurate, adaptable, and context-aware resume information extraction.

This study investigates whether integrating spatial layout features with semantic models can improve the accuracy and adaptability of resume information extraction.

Specifically, we hypothesize that a two-stage pipeline—combining OCR-based spatial recognition, section classification using XGBoost, and contextual entity extraction via Transformers—will outperform traditional methods that rely solely on textual content.

To validate this hypothesis, our research follows four main objectives:

1. Analyze existing approaches and identify their limitations,
2. Construct and annotate a dataset of resumes with spatial and semantic labels,
3. Evaluate the performance of machine learning and deep learning models for section classification and entity recognition,
4. Design and implement an integrated, hybrid information extraction pipeline.

The main contribution of this work is the development of a novel solution that combines OCR for text recognition, ML algorithms for text line classification into appropriate sections, and semantic models based on Named Entity Recognition (NER) for information extraction. This integrated approach addresses both the visual-spatial aspects of resumes and their semantic content, providing more accurate and comprehensive information extraction than current systems.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology, including the system architecture, dataset preparation, feature

engineering, and algorithms employed. Section 3 presents the experimental results, including classification and entity recognition performance. Section 4 provides a discussion of the results in the context of existing work, with analysis of contributing factors and identified limitations. Finally, Section 5 concludes the paper by summarizing the contributions and outlining directions for future research.

# 2 Method

## 2.1 System architecture

The proposed system employs a multi-stage pipeline approach for automated information extraction from resumes. The overall architecture, illustrated in Figure 1, consists of three main components: (1) text recognition and extraction using OCR, (2) text classification to identify resume sections, and (3) semantic information extraction from the classified text segments.

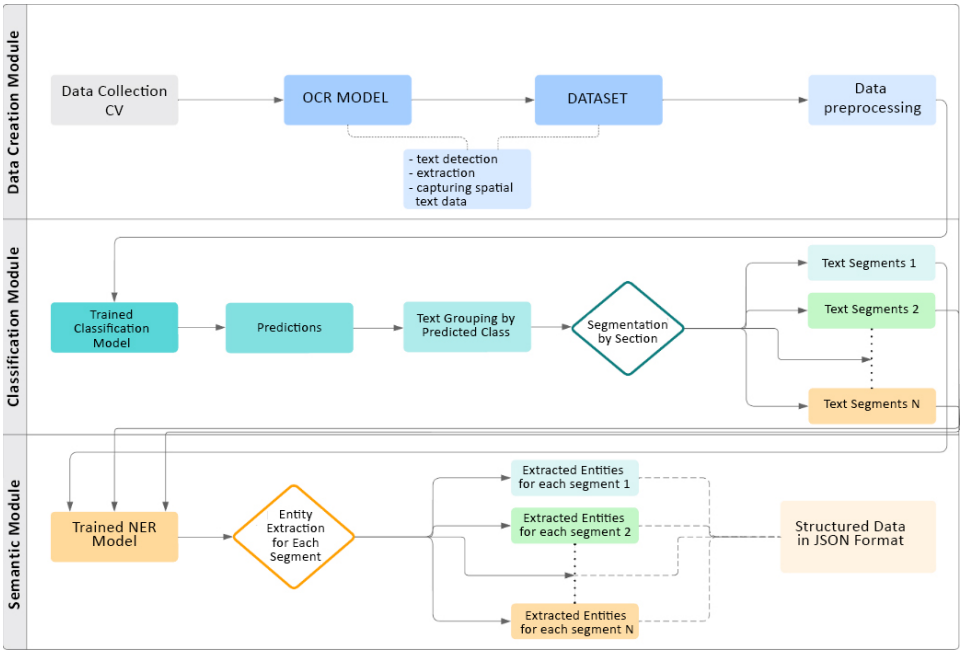



Figure 1: System architecture

The workflow begins with resume documents that are converted to images to ensure format independence. The PaddleOCR [10] model then processes these images to extract text and spatial coordinates. The extracted text lines are classified into appropriate resume sections using ML models. Finally, semantic models extract specific entities of interest from each classified section, such as candidate names, skills, education details, and work experience.

## 2.2 Dataset preparation and feature selection

### 2.2.1 Analysis of the Structure and Content of a CV

In preparing a CV, certain sections are commonly included to present relevant information for effective job applications. These sections typically include:

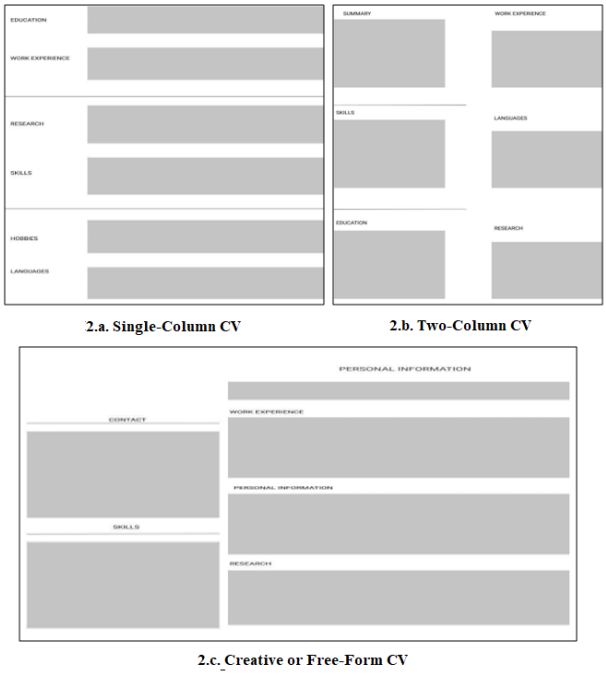
- **Personal Information:** Includes full name, address, phone numbers (home and mobile), email address, and optionally a personal website. This information allows employers to easily contact the candidate.
- **Career objective:** A short statement describing the candidate's professional goals and the type of position sought. This helps employers understand the candidate's motivations and expectations.
- **Education:** Lists academic background, including institutions attended, their locations, degrees obtained, and any relevant certifications or training.
- **Job-related skills:** Highlights specific skills relevant to the target job, whether acquired through work, internships, volunteer activities, or hobbies.
- **Professional experience:** Provides details on the candidate's work history, including company names, job titles, locations, dates of employment, and descriptions of roles and responsibilities. Relevant internships and volunteer experiences may also be included.
- **Additional information:** Covers elements that support the application such as language proficiency, computer skills, professional certifications, memberships in professional organizations, awards, and achievements. A portfolio may also be referenced if applicable.

- **Interests and activities:** Includes hobbies and leisure activities that reveal aspects of the candidate's personality and can highlight soft skills or additional qualifications.

There are four main types of CV formats, each designed to emphasize different aspects of a candidate's profile:

- **Chronological CV:** Lists the candidate's work history in reverse chronological order, starting with the most recent position. This is the most commonly used format and suits candidates with consistent career progression.
- **Functional CV:** Focuses on skills and competencies rather than the sequence of jobs. Ideal for candidates changing careers or with employment gaps, this format categorizes skills and highlights accomplishments over job history.
- **Targeted CV:** Tailored to a specific job by emphasizing the qualifications that best match the employer's expectations. It requires the candidate to carefully analyze the job posting and customize each CV section accordingly. **Combination CV:** Merges the chronological and functional formats. It begins with a summary of key competencies followed by a detailed chronological work history. This format suits candidates with both strong experience and specialized skills.

CVs can be presented in several visual formats. Figure 2 illustrates three common layout styles:




2.a. Single-Column CV 2.b. Two-Column CV

2.c. Creative or Free-Form CV

Figure 2: Common CV layout formats

- **Single-Column CV (Figure 2.a):** A traditional layout where sections are arranged vertically from top to bottom. It offers clarity and simplicity, making it easy for recruiters to read through the information.
- **Two-Column CV (Figure 2.b):** Divides the page into two main areas. The left column typically contains personal details and key skills, while the right includes professional experience, education, and other supporting content. This format improves information organization and visual balance.
- **Creative or Free-Form CV (Figure 2.c):** Often used in artistic or design-related fields, this format allows for greater customization, including asymmetric columns, infographics, colored blocks, or icons. It provides a personalized and visually distinctive presentation of qualifications.

Each of these formats offers unique advantages depending on the candidate's profile and the industry expectations.

### 2.2.2 Dataset construction for classification models

A dataset of 200 French resumes in PDF format was collected from the HR department of Intelcia IT Solutions [11]. Each resume was converted to image format to facilitate consistent processing across different layouts and styles. The PaddleOCR model was applied to extract both textual content and spatial information of each text line.

Feature engineering focused on capturing the spatial relationships between text lines and section headings within resumes. Two key types of features were developed:

1 Distance-based features: Normalized horizontal and vertical Euclidean distances between each text line and section headings were calculated. For text lines and section headings on different

pages, a specialized distance calculation was implemented that accounted for page breaks.

2 Positional features: Binary features indicating whether a text line appeared above or below each section heading were created and encoded using LabelEncoder.

3 The dataset was manually labeled with ten classes: nine representing common resume sections (Experience, Education, Skills, Projects, Certification, Languages, Interests, Software, and Personality) and a tenth class "Other" for text not belonging to any standard section. In total, 10,000 text lines were labeled to create the training corpus.

### 2.2.3 Dataset preparation for semantic models

For the semantic extraction task, text lines were grouped according to their predicted section classifications to provide contextual information. The Doccano annotation tool was used to manually annotate named entities within each section. A total of eight entity types were defined for annotation: Name, Email, Phone, Education, Experience, Skills, Language, and Certification. These categories were selected based on relevance to recruitment use cases and availability across most CVs in the dataset. The annotated text was then processed and converted to the JSONL format required by SpaCy [12] for NER model training.

## 2.3 Key algorithms

### 2.3.1 XGBoost for section classification

The eXtreme Gradient Boosting (XGBoost) algorithm was selected for resume section classification based on its superior performance. XGBoost is an ensemble learning method that builds sequential decision trees to minimize residual errors. It excels at capturing complex feature interactions and handling non-linear relationships [13].

The model was configured with the following hyperparameters:

- Maximum tree depth: 3
- Number of estimators: 100
- Learning rate: 0.1

This hyperparameter implementation allowed the model to balance complexity and generalization, as well as better capture the learning capabilities of the spatial features. XGBoost was also implemented very well in terms of its ability to mitigate model limitations from previous classification models we attempted in the study.

### 2.3.2 Artificial Neural Network for section classification

The Artificial Neural Network (ANN) was implemented as a multilayer perceptron with the following architecture:

- Input layer: Matching the dimensionality of the feature set
- Hidden layers: Two hidden layers with 64 and 32 neurons respectively
- Activation function: ReLU for hidden layers and Softmax for output layer
- Output layer: 10 neurons corresponding to the resume section classes

The model was configured with the following hyperparameters:

- Optimizer: Adam with learning rate of 0.001
- Loss function: Categorical cross-entropy
- Batch size: 32
- Training epochs: 50
- Early stopping: Patience of 5 epochs monitoring validation loss

ANNs were selected for comparison due to their proven effectiveness in text classification tasks and ability to learn complex non-linear relationships between features [14].

XGBoost was selected due to its proven performance in similar structured classification tasks. It offers efficient handling of sparse and imbalanced data, robust regularization, and interpretable feature contributions. As demonstrated later in Section 3, XGBoost outperformed alternatives such as Random Forest, ANN, and SVM, confirming its suitability for the classification of OCR-extracted resume sections.

### 2.3.3 Support Vector Machine for section classification

The Support Vector Machine (SVM) model was implemented with the following configuration:

- Kernel: Radial Basis Function (RBF)
- C parameter (regularization): 10
- Gamma parameter: 0.01
- Decision function: One-vs-Rest for multi-class classification
- Probability estimates: Enabled

SVMs were chosen for comparison due to their traditionally strong performance in text classification tasks with moderate-sized datasets and their effectiveness with high-dimensional feature spaces. The RBF kernel was selected after preliminary testing showed superior performance over linear and polynomial kernels for capturing the complex relationships in the spatial and positional features [15].

These implementations were evaluated using the same train-test split and evaluation metrics as the XGBoost model to ensure a fair comparison of performance across all three classification approaches.

### 2.3.4 Transformers model for named entity recognition

For the semantic information extraction component, a Transformers-based model was implemented using SpaCy's framework. The overall workflow for semantic model construction is illustrated in Figure 3.

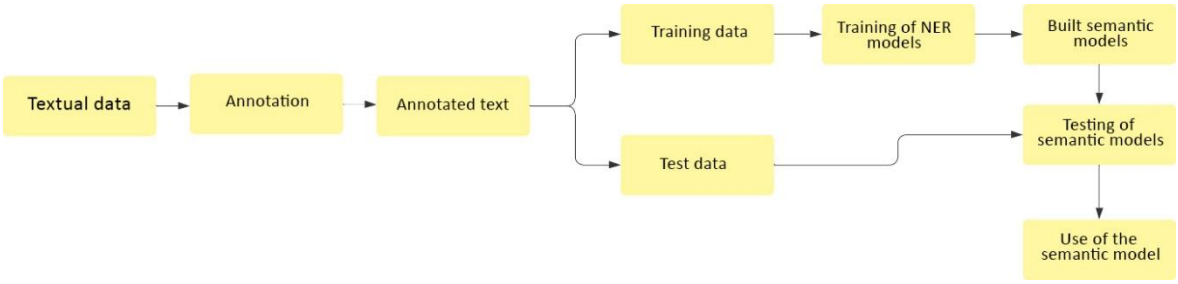



Figure 3: General workflow for semantic model construction

Transformers use an attention mechanism to capture contextual relationships between words in text sequences [16]. The semantic extraction model was built using the CamemBERT-based Transformer model, implemented through SpaCy v3.5 using the fr_dep_news_trf pipeline. CamemBERT is pretrained on large-scale French-language datasets (including OSCAR and CCNet) and employs a SentencePiece tokenizer. This choice ensured linguistic compatibility with the French resume dataset used for training. The model was fine-tuned on eight entity categories (Name, Email, Phone, Education, Experience, Skills, Language, and Certification) for 80 epochs, using the Adam optimizer and a warm-up learning rate schedule with early stopping enabled.

Training was conducted on a standard GPU environment available using Google Colab, with an average epoch runtime of 4 minutes and a total training duration of approximately 5.5 hours. The final model was exported in SpaCy's DocBin format for deployment.

The workflow begins with the classified text segments from the previous stage, which are then processed for annotation. After manual annotation using Doccano, the annotated text data is preprocessed and structured into the required format for model training. The model is then trained using the prepared dataset and evaluated against test data before final deployment.

The model was configured using a base configuration file that defined:

- Architecture parameters
- Training hyperparameters
- Optimizer settings
- Feature extraction components

The Transformers model was selected because of its ability to capture long-distance dependencies and contextual information, which is particularly valuable for identifying named entities in resume text where formatting and context provide important cues.

## 2.4 Evaluation metrics

Performance evaluation for both classification and NER models was conducted using standard metrics for multi-class classification problems [17]. First, the basic metrics for a single class are defined in equations 1 to 4:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = \frac{2 \times Recall\ \times\ Precision}{Recall + Precision} \quad (4)$$

Where:

- $TP$: True Positive, the number of cases where the model correctly predicts a positive class
- $TN$ : True Negative, the number of cases where the model correctly predicts a negative class
- $FP$ : False Positive, the number of cases where the model incorrectly predicts a positive class
- $FN$ : False Negative, the number of cases where the model incorrectly predicts a negative class

Then, for the multi-class evaluation in this study, macro-averaging was employed, which calculates in the equations 5 and 6 the metric independently for each class and then takes the average. This approach gives equal weight to all classes regardless of their frequency in the dataset:

$$Precision_{macro\_average} = \frac{1}{n}\sum_{k=1}^{n} P_k \quad (5)$$

$$Recall_{macro\_average} = \frac{1}{n}\sum_{k=1}^{n} R_k \quad (6)$$

Where $P_k$ is the precision for class $k$, $R_k$is the recall for class $k$, and $n$ the total number of classes. This evaluation ensures that performance on less frequent resume sections was properly assessed [18].

## 2.5 Pipeline overview – pseudocode

The full hybrid workflow is summarized below to illustrate the integration of the components described above.

**Algorithm 1. Hybrid Resume Parsing Pipeline**

**Input:** resume_dataset – a collection of resumes in PDF format

**Output:** Structured data with section labels and extracted named entities

```
for each resume in resume_dataset do
    image ← convert_to_image(resume)
    ocr_output ← apply_PaddleOCR(image)
    lines_with_coords← extract_text_lines_with_positions(ocr_output)
    classified_sections ← XGBoost_classify(lines_with_coords)
    for each section in classified_sections do
        ner_entities ← CamemBERT_NER(section.text)
        store(section.label, ner_entities)
    end for
end for
```

# 3 Experimental results

## 3.1 Performance comparison of classification models

The evaluation of the three classification models (ANN, SVM, and XGBoost) was conducted using a test dataset comprising 20% of the labeled data. Table 2 and figure 3 present a comparative analysis of their performance based on the evaluation metrics.

Table 2: Performance comparison of ANN, SVM, and XGBoost Algorithms

| Model | Accuracy | Macro-average precision | Macro-average recall | Macro-average F1-Score |
|---|---|---|---|---|
| ANN | 80.7% | 65.7% | 77.2% | 71.0% |
| SVM | 72.5% | 51.8% | 66.7% | 58.3% |
| XGBoost | 96.5% | 94.7% | 95.3% | 95.0% |

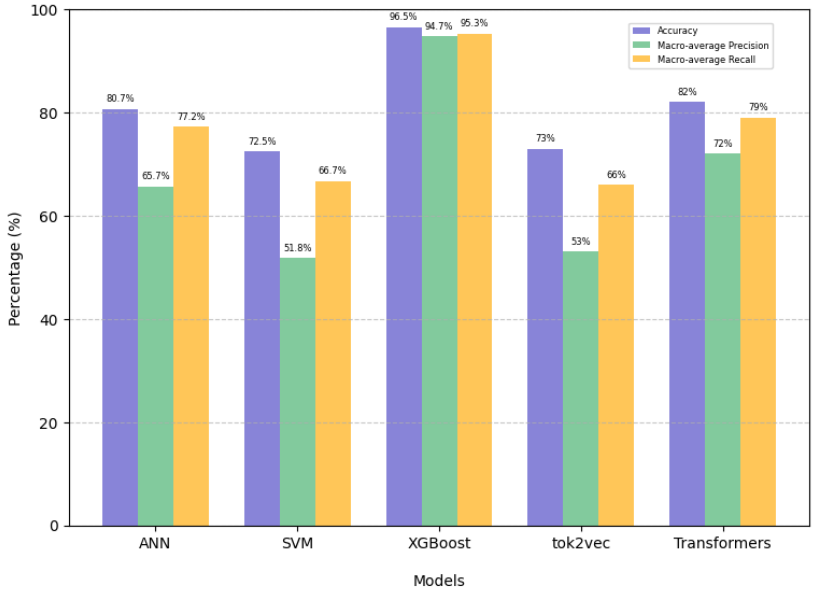



Figure 4: Performance comparison of classification and NER Models

As evident from Table 2 and figure 4, XGBoost significantly outperformed the other models across all metrics. The model achieved an impressive accuracy of 96.5%, indicating its superior ability to correctly classify text lines into their respective CV sections. Furthermore, the high macro-average precision (94.7%) and recall (95.3%) values show XGBoost's robust performance across all classes, including minority classes.

## 3.2 Performance of semantic models for named entity recognition

Two NER models were evaluated for their effectiveness in extracting named entities from the classified text: tok2vec and Transformers. Table 3 summarizes their performance after 80 training epochs.

Table 3: Comparison of NER Models: tok2vec and Transformers

| Model | Accuracy | Macro-average precision | Macro-average recall | Macro-average F1-Score |
|---|---|---|---|---|
| tok2vec | 73% | 53% | 66% | 58.7% |
| Transformers | 82% | 72% | 79% | 75.3% |

The Transformers model beat the tok2vec on the evaluation metrics overall. With an accuracy of 82%, the Transformers model was more accurate when classifying named entities in resume text.

## 3.3 Analysis of XGBoost's superior performance

XGBoost's better performance can be attributed to several factors related the nature of the algorithm:

- **Boosting technique:** XGBoost is based on a gradient boosting method that sequentially builds

new models to correct the mistakes of prior models. XGBoost is able to learn from previously misclassified items and iteratively improves prediction performance.

- **Handling complex data:** XGBoost can fit complex relationships between features, and moreover, can capture non-linear relationships. This is significant for resume texts where the spatial relationship between the text lines and section headings influences their classification.
- **Feature importance analysis:** The algorithm, in its own fashion, defines the most useful features, and classifier performance improves by emphasizing the most important features.
- **Regularization techniques**: It is noteworthy to say that XGBoost uses regularisation parameters that can assist with the likelihood of overfitting, adding to the good performance of the model on unseen data.

Figure 5 shows the confusion matrix for the XGBoost model which is indicative of its overall, high classification performance across all CV sections.

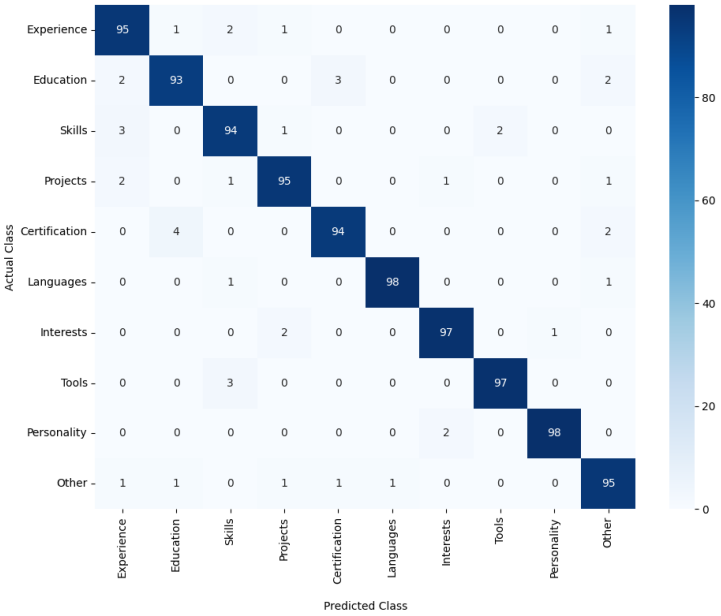



Figure 5: Confusion matrix for XGBoost Model (values in %)

The matrix shows a very high level of accuracy (93-98%) with virtually no confusion across sections where error estimates like Experience and Projects only had a 1-2% estimating error, showcasing how well the XGBoost method was able to handle the complexity of resume data.

## 3.4 Transformers model performance analysis

The superior performance of the Transformers models generally in NER task has many reasons:

- **Attention mechanism:** The Transformers model uses an attention mechanism that enables it to model contextual relationship between words. It can inspect words within a larger context where it is appearing, which enhances the accuracy of entity recognition.
- **Contextual understanding:** Rather than only focusing on the local patterns in the word sequences like in the tok2vec model, the Transformers model can also model the long-distance dependency between the words to get a more comprehensive understanding of all context in the text.
- **Sequential processing advantage:** The classification of individual lines of text accurately positioned the model to better achieve entity extraction in the Transformers module with better understanding of context.

## 3.5 Significance of the two-stage approach

An important finding of this research is the utility of the two-stage information extraction process:

1. The first stage incorporates XGBoost to classify each of the text lines into its respective CV section to help clarify for semantic analysis.
2. The second stage incorporates the Transformers model, which analyses the semantic meaning and extracts the relevant entities.

This process provides a solution to one of the main challenges of resume parsing, which was the distribution of the spatially located text within the image. By incorporating the organization of the text into sections before doing a semantic analysis, we are able to achieve a higher degree of accuracy with information extraction than purely undertaking a semantic analysis of the CV.

The results also indicate that the semantics of translating visually oriented information into semantic information creates an additional language processing dimension that

goes beyond a one-dimensional text analysis and includes both visual information and spatial language processing.

# 4 Discussion

The results presented in the previous section demonstrate that the hybrid pipeline outperforms traditional resume parsing approaches in terms of accuracy, generalization, and contextual understanding. Specifically, the XGBoost classifier achieved a 96.5% accuracy in section classification, and the Transformers model reached 82% accuracy in named entity recognition.

When compared to prior studies summarized in Table 1:

- Methods relying on rule-based or keyword techniques [1], [2], [4] showed limited adaptability to diverse resume formats and lacked semantic depth.
- Machine Learning-only approaches such as [3] achieved moderate performance (~85%) but did not incorporate spatial features or layout context.
- Deep Learning models in [7], although promising (~90%), still treated resumes as flat text, without segment-level classification or layout awareness.

In contrast, the proposed pipeline integrates both spatial (layout-aware) features and semantic (contextual) representations, which contributes to improved classification and entity recognition. The two-stage design ensures that the semantic model receives pre-structured input, enhancing its ability to extract relevant entities with higher precision.

The superior performance of the XGBoost model can be attributed to:

- Fine-grained spatial features (e.g., distances, relative positions),
- Strong regularization and ensemble learning characteristics,
- Efficient handling of imbalanced or non-linear class boundaries.

Likewise, the use of Transformers for NER offers advantages in:

- Capturing long-range dependencies across lines within the same section,
- Handling resume-specific terminology through contextual embeddings,
- Generalizing well across structurally diverse documents.

Some failure cases were observed in:

- Highly unstructured or creative resume formats (e.g., asymmetric layouts),
- Multilingual resumes, where OCR and entity recognition performance dropped,
- Misclassification between "Projects" and "Experience" when boundaries were unclear.

These cases highlight potential improvements through layout-aware Transformers or multimodal embeddings that fuse visual and textual signals.

# 5 Conclusion

This research introduces a novel hybrid AI solution for automated resume information extraction, combining OCR with Machine Learning for text classification (achieving 96.5% accuracy with XGBoost) and Deep Learning for semantic understanding (reaching 82% accuracy with Transformers). The approach addresses the challenge of resumes as spatially distributed text, where both layout and content provide crucial semantic context, demonstrating that considering spatial positioning enhances resume parsing accuracy.

While the current implementation faces limitations including language dependency, sensitivity to extreme formatting variations, and substantial training data requirements, several promising research directions emerge. Future work should explore deeper integration of visual and semantic elements, extend the approach to multi-dimensional text analysis beyond traditional linear processing, and investigate techniques requiring less labeled training data. This research ultimately points toward a new domain of natural language processing that incorporates spatially-oriented language understanding with applications extending beyond resume parsing to other complex document types.

# References

[1] Kessler, R., Torres-Moreno, J. M., & El-Bèze, M. 2010. E-Gen: automatic processing of human resources information. Document numérique, 13(3), 95–119.

[2] Baudoin, E., Déroulède, B., Diné, S., Dubouloz, M.-A., & Peretti, J.-M. 2019. Digital recruitment. In Digital transformation of the HR function (pp. 49–101). Paris: Dunod.

[3] Khan, N., Khan, K., Naveed, S., Nabi, N., Qureshi, M., & Naveed, N. 2023. Resume Parser and Summarizer. International Journal of Advanced Research in Science, Communication and Technology, 3(1), 35–42.

[4] Olorunshola, O. E., Ampitan, I. O., Adamu-Fika, F., & Ademuwagun, A. K. (2025). An Enhanced K-NN Algorithm Leveraging BERT Techniques for Resume Parsing System. Asian Journal of Research in Computer Science, 18(7), 49-59.

[5] Aakankshu, R., Kariya, J., Khant, D., Khandare, S., & Barve, P. 2020. A Systematic Literature Review (SLR) on the beginning of resume parsing in HR Recruitment Process & SMART advancements in chronological order. Research Square. https://assets.researchsquare.com/files/rs-570370/v1/9da1a6e1-437f-4f6d-a021-743ea3ee268e.pdf

[6] Gomathy, C. K. 2022. OPTICAL CHARACTER RECOGNITION. ResearchGate. https://www.researchgate.net/publication/360620085_OPTICAL_CHARACTER_RECOGNITION

[7] Sarhan, A. M., Ali, H. A., Wagdi, M., Ali, B., Adel, A., & Osama, R. (2024). CV Content Recognition

and Organization Framework based onYOLOv8 and Tesseract-OCR Deep Learning Models.
[8] Pokharel, P. 2022. Resume parser using NLP. ResearchGate. https://www.researchgate.net/publication/361772014_RESUME_PARSER
[9] Wosiak, A. 2021. Automated extraction of information from Polish resume documents in the IT recruitment process. Procedia Computer Science, 192, 2432–2439. https://doi.org/10.1016/j.procs.2021.09.012
[10] Malik, S., et al. 2020. XGBoost: A Deep Dive into Boosting. ResearchGate. https://www.researchgate.net/publication/339499154_XGBoost_A_Deep_Dive_into_Boosting_Introduction_Documentation
[11] Gao, S., Kotevska, O., Sorokine, A., & Christian, J. B. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, *16*(2), e0246310.
[12] Kumar, M., Chaturvedi, K. K., Sharma, A., Arora, A., Farooqi, M. S., Lal, S. B., ... & Ranjan, R. (2023). An algorithm for automatic text annotation for named entity recognition using Spacy framework. ICAR, Delhi, India, Tech. Rep.
[13] Chen, T., et al. 2015. XGBoost: extreme gradient boosting. R package version 0.4-2, 1(4), 1–4.
[14] Lee, J. Y., Dernoncourt, F., & Szolovits, P. 2017. Transfer learning for named-entity recognition with neural networks. arXiv preprint, arXiv:1705.06273, 1–5.
[15] Panja, S., Chatterjee, A., & Yasmin, G. 2018. Kernel Functions of SVM: A Comparison and Optimal Solution. In Advanced Informatics for Computing Research (pp. 88–97). Singapore: Springer. https://doi.org/10.1007/978-981-13-3140-4_9
[16] Ghaith, S. 2024. The triple attention transformer: advancing contextual coherence in transformer models. Evolutionary Intelligence, 17(5), 3723–3744.
[17] Riyanto, S., Imas, S. S., Djatna, T., & Atikah, T. D. 2023. Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. International Journal of Advanced Computer Science and Applications, 14(6).
[18] Grandini, M., Bagli, E., & Visani, G. 2020. Metrics for multi-class classification: an overview. arXiv preprint, arXiv:2008.05756.