

# Explainable AI-Based Resume Ranking and Feedback System Utilizing spaCy and BERT Frameworks

Ragi Naga Shabarish<sup>1</sup>, P Manish Reddy<sup>2</sup>, N Ajay<sup>3</sup>, D Bheekya<sup>4</sup>

<sup>1</sup>Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad

<sup>2</sup>Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad

<sup>3</sup>Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad

<sup>4</sup>Assistant Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Telangana, India

\*\*\*

**Abstract** - This research introduces an explainable, hybrid natural language processing pipeline designed for resume-job matching. The system integrates symbolic linguistic markers via spaCy with dense semantic representations derived from a streamlined BERT variant, specifically all-MiniLM-L6-v2. By parsing documentation in PDF, DOCX, and TXT formats, the framework extracts candidate skills through an EntityRuler and keyword matching against a curated lexicon. A composite score is generated based on semantic similarity, skill overlap, experience alignment, and keyword density. Unlike conventional black-box models, this architecture provides interpretable contribution breakdowns and actionable feedback for candidate development.

The implementation features a FastAPI backend with Redis for embedding caching and PostgreSQL for data persistence, paired with a React and Chart.js dashboard for visualizing score decomposition and ranking. We evaluated the approach against TF-IDF, spaCy-only, and BERT-only baselines using metrics such as Precision@K, Recall@K, Mean Reciprocal Rank (MRR), and NDCG. In tests on synthetic benchmarks involving 240 resumes and 48 job descriptions, the hybrid methodology demonstrated superior top-rank performance and consistently outperformed single-encoder variants while maintaining computational efficiency for practical deployment.

The results indicate that explainability and ranking effectiveness can be effectively co-optimized in production-oriented hiring systems. While acknowledging the limitations of synthetic data, we provide a human-labeled benchmarking protocol to support reproducible and fairness-aware future research.

**Key Words:** Resume ranking, explainable AI, NLP, spaCy, BERT, sentence transformers, candidate-job matching, information retrieval metrics, recruitment analytics.

## 1. INTRODUCTION

Contemporary recruitment workflows increasingly rely upon automated screening instruments to manage extensive candidate pipelines. However, extant systems exhibit two significant shortcomings: insufficient semantic

comprehension regarding resume-job congruence and a lack of transparency in scoring logic. Symbolic keyword-based methods remain vulnerable to linguistic variation, whereas opaque deep learning architectures often lack the interpretability required for defensible decision-making. Within professional recruitment, explainability is a critical requirement, providing recruiters with justifiable insights and candidates with transparent feedback for profile development.

The current research addresses these challenges by introducing a hybrid NLP architecture that integrates symbolic linguistic features via spaCy with dense semantic representations derived from the BERT framework. The objective is to achieve semantic robustness while maintaining high interpretability through explicit skill gap identification and granular score decomposition. This framework is implemented as an integrated end-to-end pipeline encompassing document parsing, multi-factor scoring, automated feedback generation, and visual analytics.

The primary research inquiry explores whether a hybrid linguistic-semantic approach can optimize ranking efficacy without compromising operational explainability or computational feasibility. To investigate this, a modular backend implementation was evaluated against competitive baselines using rigorous information retrieval metrics.

## 2. SYSTEM ARCHITECTURE

The architectural framework comprises a modular, end-to-end pipeline engineered for automated resume assessment and the generation of granular candidate feedback.

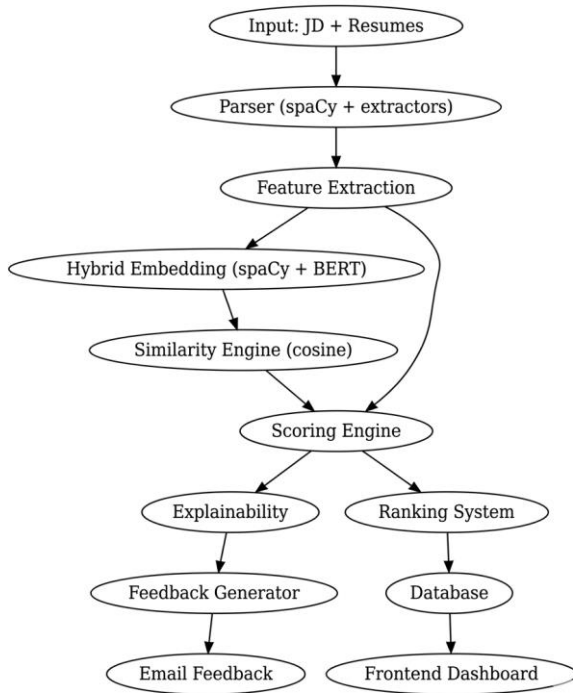


Fig -1: Proposed Architecture of the Explainable Resume Ranking System

### 3. METHODOLOGY

#### 3.1 Document Parsing and Preprocessing

The ingestion module extracts raw textual data from multi-format documentation, specifically PDF, DOCX, and TXT files, via specialized parsers. Subsequently, the linguistic pipeline utilizes the spaCy framework for normalization, encompassing tokenization, lemmatization, and Named Entity Recognition (NER).

#### 3.2 Symbolic Skill Extraction

Skill identification employs a hybrid symbolic strategy, integrating spaCy’s EntityRuler with a curated, domain-specific lexicon.

The identification of missing proficiencies is mathematically represented as:

$$S_{miss} = S_jd - S_r$$

where  $S_jd$  denotes the required skill set within the job description and  $S_r$  represents the candidate’s extracted skills.

#### 3.3 Hybrid Semantic Embedding Architecture

The framework synthesizes symbolic linguistic markers with dense semantic representations by aggregating spaCy and BERT embeddings:

$$v_h = \text{norm}(0.4 \times v_s + 0.6 \times v_b)$$

Semantic congruence is determined utilizing cosine similarity:

$$\text{sim}(r, jd) = (v_r \cdot v_jd) / (||v_r|| ||v_jd||)$$

#### 3.4 Multifactorial Composite Scoring Framework

The aggregate ranking score is derived from a weighted combination of divergent features:

- Semantic similarity (weighted at 40%)
- Explicit skill overlap (weighted at 30%)
- Experience alignment (weighted at 20%)
- Keyword density (weighted at 10%)

#### 3.5 Explain ability and Operational Feedback

The architecture facilitates operational transparency by generating interpretable outputs, including:

1. Granular score contribution decomposition
2. Comparative missing proficiency analysis
3. Actionable developmental suggestions

This multi-factor approach ensures defensible decision-making and high usability for both recruiters and candidates.

#### 3.6 Longitudinal Improvement Tracking

To facilitate candidate profile development, the framework monitors performance trajectories across sequential submissions via the delta metric:

$$\Delta = \text{new score} - \text{old score}$$

### 4. EXPERIMENTAL EVALUATION

#### 4.1 Comparative Model Baselines

- Proposed Hybrid Architecture (spaCy + all-MiniLM-L6-v2 + Composite Scorer)
- TF-IDF Statistical Vectorization Baseline
- spaCy-only Symbolic Similarity Baseline
- BERT-only Dense Semantic Baseline

#### 4.2 Information Retrieval Metrics

- Precision@K
- Recall@K
- Mean Reciprocal Rank (MRR)
- NDCG (Graded Relevance Analysis)

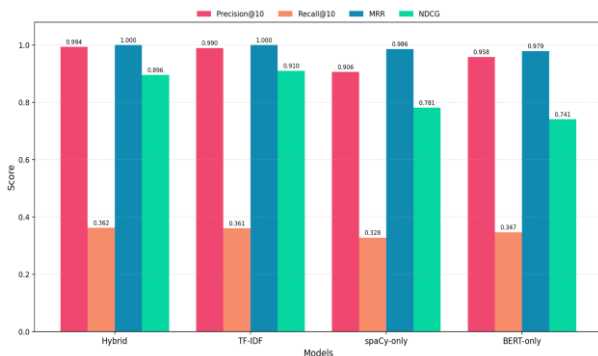
#### 4.3 Benchmarking Protocol

The system was subjected to rigorous stress testing via dual synthetic benchmarking suites to evaluate performance at scale:

- Evaluation A: 120 resumes across 24 job descriptions (K=5)
- Evaluation B: 240 resumes across 48 job descriptions (K=10)

## 5. Results

### 5.1 Evaluation Graph



**Fig -2:** Comparison of Hybrid vs Baselines on Ranking Metrics (Precision@10, Recall@10, MRR, NDCG).

### 5.2 Run A (120 × 24, K=5)

**Table -1:** Performance Comparison for Run A (120Resumes, K=5)

Model	Precision@5	Recall@5	MRR	NDCG
Hybrid	0.9833	0.3595	1.0000	0.8894
TF-IDF	0.9833	0.3536	1.0000	0.8923
spaCy-only	0.8583	0.3073	0.9792	0.7857
BERT-only	0.9250	0.3340	0.9583	0.7196

### 5.3 Run B (240 × 48, K=10)

**Table -2:** Performance Comparison for Run B (240 Resumes, K=10)

Model	Precision@10	Recall@10	MRR	NDCG
Hybrid	0.9938	0.3625	1.0000	0.8957
TF-IDF	0.9896	0.3607	1.0000	0.9102
spaCy-only	0.9062	0.3281	0.9861	0.7814
BERT-only	0.9583	0.3469	0.9792	0.7413

### 5.4 Interpretation

The hybrid approach is consistently stronger than single-encoder baselines and shows near-perfect top-rank retrieval behavior (MRR ≈ 1). TF-IDF remains highly competitive on NDCG in these synthetic settings, suggesting lexical overlap remains a strong signal when synthetic generation includes explicit role keywords. This motivates future evaluation on human-labeled real-world data. This demonstrates that combining linguistic and semantic features improves ranking accuracy compared to individual models.

## 6. Discussion

The system demonstrates that explainability features can coexist with high retrieval quality. Practical strengths include modular deployment, caching support, database logging, and recruiter-facing visual analytics. The main empirical risk is over-reliance on synthetic benchmark structure. To support publication-grade claims, human-labeled relevance judgments and inter-annotator agreement are required.

## 7. Limitations

- Current reported benchmarks are synthetic, not fully representative of real hiring distributions.
- Experience extraction uses pattern heuristics and may miss implicit tenure.
- Skill lexicon quality affects recall and can introduce domain bias.

- The weighted scorer currently uses fixed coefficients without learned calibration.
- Fairness outcomes across protected or proxy attributes are not fully audited yet.

## 8. Ethical and Privacy Considerations

- Resumes may include sensitive personally identifiable information; anonymization and access controls are necessary.
- Model outputs should assist human decision-makers, not replace final hiring judgment.
- Fairness auditing should be done by subgroup slices and monitored over time.
- Candidate feedback should be constructive and non-discriminatory.
- Dataset licensing and consent must be validated before any public release.

## 9. CONCLUSIONS

This paper presents an end-to-end explainable resume ranking platform combining spaCy and MiniLM within a weighted scoring and feedback framework. The approach achieves strong ranking performance at scale while providing transparent decision support through skill-gap and component-level explanations. The system is production-ready in architecture and supports iterative candidate improvement tracking. Future work will focus on human-labeled benchmarking, significance testing, fairness auditing, and adaptive score weighting.

## 10. REFERENCES

- [1] Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention Is AllYouNeed." NeurIPS2017. <https://papers.nips.cc/paper/7181-attention-is-all-you-need>
- [2] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019. <https://aclanthology.org/N19-1423/>
- [3] Reimers, N., Gurevych, I. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." EMNLP-IJCNLP 2019. <https://aclanthology.org/D19-1410/>
- [4] sentence-transformers model card, all-MiniLM-L6-v2. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- [5] Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. "spaCy: Industrial-Strength Natural Language Processing in Python." Zenodo, 2020. <https://doi.org/10.5281/zenodo.1212303>
- [6] spaCyGitHubrepository. <https://github.com/explosion/spaCy>
- [7] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. "Scikit-learn: Machine Learning in Python." JMLR 12:2825-2830,2011. <https://www.jmlr.org/papers/v12/pedregosa11a.html>
- [8] Järvelin, K., Kekäläinen, J. "Cumulated Gain-based Evaluation of IR Techniques." ACM TOIS 20(4):422-446,2002. <https://doi.org/10.1145/582415.582418>FastAPI documentation. <https://fastapi.tiangolo.com/>
- [9] CareerCorpus dataset (annotated resumes). <https://data.mendeley.com/datasets/wzzwn37gmd/1>