**AI RESUME ANALYZER AND JOB MATCHER**

# 

# Project Documentation

Submitted in Partial fulfilment of the Requirements for the award of the Degree of

**MASTER OF SCIENCE (INFORMATION TECHNOLOGY)**

## By

**SHRUTI SUDAY HARAYAN** Seat No. **202**

# Under the esteemed guidance of Prof. Suryaprakash Upadhyay Assistant Professor



DEPARTMENT OF INFORMATION TECHNOLOGY **RAMANAND ARYA D.A.V COLLEGE, AUTONOMOUS**

## (Affiliated to University of Mumbai) MUMBAI, 400042

## MAHARASHTRA

**2025-26**

**RAMANAND ARYA D.A.V COLLEGE, AUTONOMOUS**

#### (Affiliated to University of Mumbai)

MUMBAI, MAHARASHTRA – 400042

**DEPARTMENT OF INFORMATION TECHNOLOGY**

****

**CERTIFICATE**

This is to certify that project entitled, “**AI RESUME ANALYZER AND JOB MATCHER**”, is a bonafide work of **SHRUTI SUDAY HARAYAN** bearing Seat No. **202** Submitted in the partial fulfilment of the requirements for the award of Degree of  **MASTER OF SCIENCE in INFORMATION TECHNOLOGY** from the University of Mumbai.

**Internal Guide Coordinator**

**External Examiner**

**Date : College Seal**

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who helped me complete this research successfully and I came to know about so many new things.

First and foremost, I would like to thank to **CA Dr. Ravindra Naik**, **I/C Principal of Ramanand Arya DAV college,** who gave me golden opportunity to do this wonderful project.

I would like to express my special thanks and gratitude to the project guide **Assistant** **Professor Suryaprakash Upadhyay** for guiding me to do the project work on time and giving support and guidance, which made me complete my project duly. I am extremely thankful to him for providing such nice support and guidance.

I am thankful for and fortunate enough to get constant encouragement, support, and guidance from the teachers of Information Technology who helped me in successfully completing my project work.

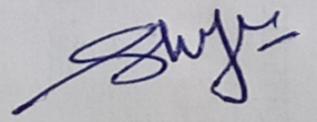
Finally, I would like to thank my family who motivated me and boosted my morale.

DECLARATION

I hereby declare that the project entitled, **“AI RESUME ANALYZER AND JOB MATCHER”** done at RAMANAND ARYA DAV COLLEGE, has not been in any case duplicated to submit to any other universities for the award of any degree. To the best of my knowledge, no one has submitted to any other university.

The project is done in partial fulfilment of the requirements for the award of degree of **MASTER OF SCIENCE (INFORMATION TECHNOLOGY)** to be submitted as final semester project as part of our curriculum.

**SHRUTI S. HARAYAN**



**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Chapter** | **Page no.** | **Sign** |
| 1 | Introduction | 6- |  |
| 2 | Description of Problem  2.1 research gaps |  |  |
| 3 | Literature Review |  |  |
| 4 | Research Methodology |  |  |
| 5 | Code Implementation |  |  |
| 6 | Conclusion |  |  |
| 7 | Reference |  |  |

**INTRODUCTION**

## ****1.1 Background****

Recruitment and hiring have always been critical processes in any organization. With the exponential rise in job applications and the increasing complexity of skill requirements, traditional manual resume screening has become inefficient, error-prone, and time-consuming. Human recruiters face challenges such as bias, inconsistency, and fatigue when evaluating large volumes of resumes for specific roles.

To address these limitations, **Artificial Intelligence (AI) powered Resume Analyzers** have emerged as innovative solutions. Such systems use Natural Language Processing (NLP), Machine Learning (ML), and Semantic Analysis to evaluate resumes beyond simple keyword matching. Unlike traditional Applicant Tracking Systems (ATS), which rely heavily on exact keyword matches, AI-based models (e.g., Sentence-BERT, Transformers) enable semantic understanding of resumes, allowing recruiters to identify relevant skills, experiences, and qualifications even if they are expressed differently than in the job description.

Given the growing interest in AI for recruitment, a research-driven **AI Resume Analyzer** can provide a deeper understanding of how intelligent automation can reduce hiring inefficiencies, enhance fairness, and improve candidate-job matching accuracy.

## ****1.2 Problem Statement****

The hiring process often faces several challenges:

* Recruiters spend excessive time manually filtering resumes, which slows down the recruitment cycle.
* Traditional ATS systems mostly rely on keyword matching, which leads to high false positives (resumes that match keywords but lack relevant experience) and false negatives (resumes with required skills but expressed differently).
* Bias in resume screening, whether conscious or unconscious, often affects diversity and fairness in hiring.
* Job seekers are frequently unaware of how their resumes are interpreted by ATS, leading to mismatches and missed opportunities.

Therefore, there is a need for an **AI-powered Resume Analyzer** that not only matches resumes with job descriptions using advanced semantic similarity techniques but also provides feedback to candidates on how to improve their resumes. This research seeks to bridge the gap between keyword-based ATS systems and context-aware, intelligent recruitment solutions.

## ****1.3 Research Objectives****

The major objectives of this study are:

* **To design and develop** an AI-powered Resume Analyzer that leverages NLP and semantic similarity models for improved resume-job matching.
* **To compare** the effectiveness of AI-based semantic matching with traditional keyword-based ATS systems.
* **To evaluate** the accuracy, precision, recall, and fairness of the AI Resume Analyzer in real-world recruitment datasets.
* **To provide** candidates with actionable insights, such as missing skills, relevant keywords, and match scores, thereby enhancing resume quality.
* **To explore** how AI-powered tools can reduce human bias and improve the overall fairness and inclusivity of recruitment.
* **To contribute** to the body of research on AI in recruitment by proposing a framework for context-aware and fair resume analysis.

## ****1.4 Motivation of the Study****

The core motivation for this research stems from the urgent need to streamline the recruitment process while upholding fairness, efficiency, and transparency. Existing manual methods and even basic applicant tracking systems (ATS) fall short of addressing these needs in a scalable manner.

By leveraging AI and ML, there is an opportunity to bridge the gap between job seekers and recruiters, making the hiring process more accessible, precise, and informative for all stakeholders.

* **Recruiter’s Motivation**:  
  Modern organizations are overwhelmed with job applications, sometimes receiving thousands of resumes for a single position. Recruiters require efficient, unbiased, and accurate systems that can streamline candidate shortlisting.
* **Candidate’s Motivation**:  
  Many skilled candidates are rejected due to poorly optimized resumes that fail to pass through ATS filters. Providing job seekers with AI-powered insights on resume improvements can help them showcase their skills more effectively and increase job opportunities.
* **Researcher’s Motivation**:  
  While many commercial ATS solutions exist, there is limited academic research on integrating **semantic similarity models** (such as Sentence-BERT, SBERT, or transformer-based encoders) into resume analysis. Conducting this study bridges the gap between theoretical research and practical applications in recruitment technology.
* **Societal Motivation**:  
  The study contributes to fair hiring practices by reducing unconscious bias and ensuring that candidates are evaluated more on their actual skills and experiences rather than superficial keyword matching.

Ultimately, this research aims to advance the state of AI-powered recruitment tools, promoting ethical, transparent, and efficient hiring practices in an increasingly digital world.

**Description of Problem**

Recruitment has become increasingly complex in the modern job market, where organizations often receive hundreds or even thousands of applications for a single job posting. Traditional manual resume screening methods are not only **time-intensive** but also highly **susceptible to human errors and biases**, making them unsustainable in large-scale hiring. To cope with the volume of applications, many companies now employ Applicant Tracking Systems (ATS). However, existing ATS solutions are limited in several critical ways.

Most ATS platforms rely on **keyword-based filtering**, where resumes are matched to job descriptions by detecting the presence of specific words or phrases. While this method offers speed, it introduces significant shortcomings:

* **Qualified candidates may be rejected** if their resumes use synonyms or alternative expressions instead of the exact keywords from the job description.
* **Unqualified candidates may be shortlisted** if they strategically include keywords, even without relevant experience.
* **Job seekers receive little to no feedback**, leaving them unaware of why their applications were unsuccessful.
* **Fairness and inclusivity remain a challenge**, as automated screening can reinforce hidden biases based on factors such as gender, educational background, or naming conventions.

These limitations result in inefficiencies for recruiters and missed opportunities for talented candidates, ultimately hindering the overall recruitment process.

Overcoming these challenges is therefore of **critical importance**. From an organizational perspective, integrating advanced AI-based techniques can improve the **accuracy, efficiency, and fairness** of resume screening, ensuring that recruiters focus on genuinely qualified candidates. For job seekers, providing **transparent feedback** and highlighting **missing skills** can empower them to improve their resumes and increase their chances of securing employment.

From a research standpoint, developing an AI Resume Analyzer that leverages **semantic similarity models (e.g., SBERT, Transformers)** contributes to the growing field of intelligent recruitment technologies, bridging the gap between academic research and industry practices.

Thus, the central problem addressed in this study is the **inefficiency, lack of semantic understanding, limited transparency, and potential bias** in existing resume screening systems, and the importance of designing an AI-powered solution that overcomes these limitations to enhance both recruiter and candidate experiences.

## ****2.1 Research Gap****

Despite the rapid advancement of recruitment technologies and the widespread adoption of Applicant Tracking Systems (ATS), several critical gaps remain unaddressed in the current literature and industry practices:

1. **Over-Reliance on Keyword Matching**

* Most ATS solutions focus on matching resumes with job descriptions based on exact keywords.
* This approach fails when candidates use synonyms, different terminologies, or domain-specific phrases to describe their skills and experiences.
* As a result, qualified candidates are often overlooked, while irrelevant candidates may pass through screening.

1. **Lack of Semantic Understanding**

* Existing systems rarely incorporate advanced Natural Language Processing (NLP) models such as Transformer-based architectures (e.g., BERT, SBERT).
* Without semantic similarity analysis, ATS cannot fully understand context, intent, and conceptual equivalence between resumes and job descriptions.

1. **Limited Feedback Mechanisms for Job Seekers**

* Most ATS act as “black boxes” — candidates are either shortlisted or rejected without understanding the reasons.
* Very few systems provide **explainability** or feedback such as missing skills, keyword recommendations, or suggestions for improvement.

1. **Bias and Fairness Concerns**

* Automated recruitment systems risk amplifying existing biases (e.g., based on gender, name, or educational background).
* There is a research gap in developing **fair, transparent, and bias-aware** AI models for resume screening.

1. **Insufficient Benchmarking and Research Validation**

* While industry solutions exist, academic research on **evaluating and benchmarking AI resume analyzers** against real-world recruitment datasets is limited.
* Few studies compare traditional keyword-based ATS performance with modern AI-driven semantic models in a systematic way.

### ****Contribution of This Study****

To address the above gaps, this research aims to:

* Incorporate **semantic similarity models** (SBERT) into resume-job matching, moving beyond simple keyword matching.
* Develop a framework that provides **explainable feedback** to candidates, enabling them to improve their resumes.
* Explore methods to **reduce algorithmic bias** and ensure fairness in recruitment processes.
* Benchmark the system against traditional ATS to demonstrate improvements in **accuracy, recall, and fairness**.

Thus, this study contributes to both **academic research** and **practical recruitment solutions** by proposing an intelligent, fair, and feedback-driven AI Resume Analyzer.

**Literature Review**

## 2.1 Introduction to the Literature Review

Recruitment technology has shifted from manual screening and keyword-driven Applicant Tracking Systems (ATS) toward **semantic, context-aware NLP**. Recent works show that **Transformer-based embeddings** (e.g., SBERT) substantially improve resume–JD alignment versus lexical matching alone, while regulators increasingly scrutinize fairness and transparency in AI-assisted hiring. This review synthesizes (i) recruitment/ATS evolution, (ii) AI/NLP approaches for resume–JD matching, (iii) semantic similarity advances, (iv) bias/fairness and regulation, and (v) candidate-facing explainability and feedback—then critically analyzes gaps my system targets.

## 2.2 Thematic Review of Prior Work

### 2.2.1 Recruitment and Resume Screening (From Keywords to Semantics)

Early ATS tools emphasized **lexical filters** (exact tokens, Boolean search). While scalable, they miss qualified candidates using synonyms and are vulnerable to “**keyword stuffing**,” where hidden or repeated terms inflate scores without true fit. Industry and journalism have documented these failures, motivating semantic alternatives.

Recent academic and practitioner reports argue configuration choices in ATS can **reject capable candidates** and narrow talent pools, signaling the limits of one-dimensional keyword pipelines.

### 2.2.2 AI and NLP in Recruitment

ML/NLP methods automate parsing, skill extraction, and matching. Surveys and empirical studies outline pipelines that tokenize resumes/JDs, extract entities/skills, and compute similarity scores for ranking or recommendation. Comparative works show AI-based screening increases throughput but surface **new concerns** around transparency and bias.

Recent research prototypes explore **embedding-based job–resume matching** (e.g., Resume2Vec; vector search approaches), often outperforming classical baselines in human-aligned metrics and ranking quality.

### 2.2.3 Semantic Similarity and Transformer-based Models

**BERT** introduced deep bidirectional language representations that improved sentence-pair tasks but were costly for large-scale retrieval. **Sentence-BERT (SBERT)** adapted BERT into a **Siamese architecture** that yields sentence embeddings suitable for **cosine-similarity** search—enabling efficient, accurate semantic matching at scale. Numerous resume-JD studies now employ **SBERT + cosine.**

Cutting-edge, domain-specific variants for labor data have appeared, e.g., **CareerBERT** (resume–ESCO job matching), and **SkillMatch** (self-supervised skill relatedness benchmark and SBERT adaptation to job-ad skill co-occurrence), underscoring the field’s shift to **skill-centric, semantic modeling**.

### 2.2.4 Hybrid Scoring: Combining Semantics and Keywords

Several IR and applied AI sources promote **hybrid search** that blends **keyword signals** with **semantic similarity** to improve precision/recall trade-offs. This hybridization mitigates false positives from embeddings alone and false negatives from sparse lexical overlap. The **strictness penalty** (down-weighting semantic similarity when overlap=0) is a pragmatic defense against generic language and adversarial keyword stuffing documented in practice.

Empirical works on resume matching increasingly report **cosine-based embedding scores** complemented by auxiliary features (skills overlap, section weights). Recent comparisons (e.g., BERT vs. LLM-rankers) also highlight **context length and explainability trade-offs**, supporting the case for stable, interpretable hybrids in production-like pipelines.

### 2.2.5 Fairness, Bias, and Regulatory Guidance

Regulators stress that AI hiring tools must **avoid disparate impact** and remain explainable. The U.S. **EEOC** launched its **AI & Algorithmic Fairness Initiative** and published guidance spotlighting risks of discrimination when automated tools influence selection. Ongoing legal scrutiny (e.g., suits involving hiring software) illustrates the compliance stakes for resume-screening systems.

Academic studies show AI recommendations can **shift human decisions**—sometimes mitigating, sometimes introducing different biases—hence the need for **transparent reasoning and bias controls** in recommender interfaces.

### 2.2.6 Candidate Experience, Transparency, and Feedback

User studies indicate applicants often **welcome AI** when it is perceived as useful and easy to use, but trust hinges on **explanations** and **actionable feedback**. Systems that disclose why a score was assigned and how to improve (e.g., missing skills, ATS-unfriendly formatting) bolster acceptance and perceived fairness.

## 2.3 Critical Analysis of Prior Studies

* **Strengths of Semantic Models:** SBERT-style embeddings deliver robust **contextual similarity** with tractable retrieval time, consistently outperforming bag-of-words and TF-IDF on sentence-level relevance and clustering. However, pure embeddings may over-score resumes sharing **generic language** with JDs (e.g., “team player,” “fast-paced environment”), justifying **keyword/skill constraints** as a counterbalance.
* **Limits of Keyword-Only ATS:** Lexical filters are brittle to synonyms and format variance, and are exploitable via **keyword stuffing**. Research and reportage show configuration pitfalls can silently **filter out qualified candidates**, motivating hybrid or semantic-first scoring and **format checks** (e.g., tables/graphics warnings).
* **Explainability & Feedback:** Many academic systems report accuracy gains but provide **limited candidate-facing explanations**. Inclusion of **reason codes**, **matched/missing skills**, and **tone-aware recommendations** directly addresses this gap and aligns with adoption research findings.
* **Fairness, Compliance, and Evaluation:** A growing body of guidance urges **bias testing** and documentation for AI hiring tools. Yet, benchmark datasets with **demographic annotations** and standardized fairness metrics for resume matching remain scarce, complicating apples-to-apples comparisons.
* **Latest Research Directions:** Domain-adapted models (CareerBERT), **hypothetical/contrastive finetuning** for sparse labels (ConFit v2), and **vector-search pipelines** indicate momentum toward scalable, skill-aware matching—suggesting future gains for systems, especially if they fine-tune on **HR-domain skill graphs** and adopt **hybrid retrieval**.

## 2.4 How This Research Method Fits the Literature

* **Semantic core:** **SBERT cosine similarity** between cleaned resume and JD → aligns with best-practice semantic matching and with multiple recent studies reporting strong retrieval/ranking performance.
* **Keyword/skill signal:** **Normalized skill overlap** using a master skills list + **aliases** → mirrors hybrid IR guidance; directly combats synonym gaps and improves precision.
* **Hybrid scoring:** **Fixed 0.5/0.5 weights** for semantic vs. keywords → consistent, reproducible ranking (and easily tunable via validation). Hybridization is widely recommended to balance recall and precision.
* **Strictness penalty when overlap=0:** Penalizing resumes with **zero skill overlap** prevents generic SBERT similarity from inflating scores—an effective defense consistent with literature noting adversarial keyword behaviors and generic-language pitfalls.
* **Explainability:** explain\_ats\_score() and recruiter-style get\_recommendations() address adoption and trust concerns surfaced in applicant-perception research.
* **ATS-format checks:** Warnings for **tables/graphics/HTML** reflect known parsing issues in legacy pipelines and help candidates pass automated filters.

# ****Evaluation and Future Work****

## ****1. Evaluation Framework****

### 1.1 Evaluation Metrics

The effectiveness of the proposed AI Resume Analyzer is measured through ranking and classification metrics widely used in information retrieval and recommendation systems:

**Area Under the ROC Curve (AUC):**  
Evaluates the system’s ability to distinguish between relevant and non-relevant resumes for a given job description. A higher AUC value indicates stronger discrimination performance.

**Precision@k:**  
Measures the proportion of relevant resumes among the top-k ranked candidates. For example, Precision@5 indicates the percentage of correct matches within the top five resumes retrieved.

**Normalized Discounted Cumulative Gain (nDCG):**  
Captures the quality of ranking by rewarding systems that place more relevant resumes higher in the order. nDCG accounts for both the relevance of a candidate and their position in the ranking.

Together, these metrics provide a holistic view of performance by measuring both **accuracy** (AUC) and **ranking quality** (Precision@k, nDCG).

### 1.2 Ablation Study

To validate the contributions of different components in the proposed hybrid model, an ablation study is recommended. This involves comparing variants of the system:

**SBERT-only:** Matching based solely on semantic embeddings.

**Keyword-only:** Matching based solely on keyword overlaps.

**Hybrid:** Combination of semantic and keyword approaches.

**Hybrid + Strictness Penalty:** Final version, which incorporates penalties for irrelevant keyword inflation.

Such experiments demonstrate the incremental value of each module and strengthen the empirical justification for the proposed design.

## ****2. Future Work and Extensions****

### 2.1 Fairness and Ethical Considerations

Automated resume screening systems must be evaluated for potential bias and fairness to ensure compliance with legal and ethical standards. According to the **U.S. Equal Employment Opportunity Commission (EEOC)**, hiring systems should avoid adverse impact on protected demographic groups.

**Adverse Impact Ratio (4/5th Rule):**  
Selection rates for any demographic group should not be less than 80% of the highest group’s selection rate. For example, if 60% of male candidates are shortlisted, at least 48% of female candidates should also be shortlisted.

**Fairness Probes:**  
Synthetic or vignette datasets can be constructed to evaluate whether the system disproportionately disadvantages certain groups. Toolkits such as **AI Fairness 360 (IBM)** or **FairLearn (Microsoft)** can be integrated for fairness auditing.

Addressing fairness not only improves regulatory compliance but also increases organizational trust in AI-powered recruitment.

### 2.2 Explainability and Transparency

Recruitment decisions have high stakes; hence, users should be able to understand **why** a resume was matched or rejected. Future extensions may include:

Highlighting matched skills and keywords directly in the resume.

Providing natural language explanations (e.g., “Candidate was ranked highly due to expertise in Java, Spring Boot, and cloud technologies”).

Integrating **XAI (Explainable AI)** methods such as SHAP or LIME to interpret model behavior.

### 2.3 Domain Adaptation and Skill Graph Integration

General-purpose semantic models such as SBERT may not capture domain-specific occupational structures. To improve contextual relevance:

**Skill-Aware Fine-Tuning:** Adapt embedding models using structured skill ontologies such as **ESCO (European Skills, Competences, and Occupations)** or **O\*NET (U.S. Occupational Information Network)**.

**Domain Adaptation:** Train or fine-tune models on industry-specific datasets (e.g., IT resumes vs. Healthcare resumes).

**SkillMatch-style Approaches:** Recent research suggests aligning candidate-job embeddings with external labor market skill graphs for higher accuracy and transferability.

Such enhancements can increase robustness across sectors, improve adaptability to labor market shifts, and enhance recruiter trust in system recommendations.

## ****Conclusion****

The current evaluation framework provides a baseline for measuring system effectiveness through standard ranking metrics and ablation analysis. However, for the AI Resume Analyzer to become a production-ready solution, it must integrate fairness audits, explainability features, and domain adaptation strategies. These extensions not only address ethical and regulatory considerations but also ensure the system remains scalable, reliable, and equitable in real-world hiring contexts.

## 2.5 Research Gaps Emerging from the Literature

* **Standardized, open benchmarks** for resume–JD matching with **human judgments** and **fairness labels** are limited, hampering rigorous comparison.
* **Explainable scoring frameworks** that surface **actionable, candidate-level feedback** (not just global model interpretations) are still rare in peer-reviewed studies.
* **Robustness to adversarial formatting/phrasing** (keyword stuffing, resume templates) remains under-studied; strictness strategies warrant formal evaluation.
* **Regulatory-aligned evaluations** (documentation, bias audits) need to be mainstreamed into technical papers to meet compliance expectations.

## 2.6 Summary and Implications for This Study

The literature shows a clear transition from **keyword-centric ATS** to **semantic, embedding-based** matching, with **hybrid scoring** emerging as a practical best practice. This project operationalizes these advances: **SBERT + cosine** for context, **skill overlap** for precision and transparency, and a **strictness penalty** to guard against generic matches. What remains is to **empirically validate** the design on real or semi-synthetic datasets, incorporate **explainability and fairness** measures aligned with **regulatory guidance**, and consider domain adaptation to labor-market skill graphs.

## 2.7 Representative Recent Works (2022–2025)

**SBERT & semantic similarity:** Reimers & Gurevych (2019) — foundational; widely adopted in hiring research. ([arXiv](https://arxiv.org/abs/1908.10084?utm_source=chatgpt.com" \o "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks))

**Enhanced resume screening with SBERT:** Fast cosine-based pipelines showing high screening accuracy and throughput. ([The Science and Information Organization](https://thesai.org/Downloads/Volume15No8/Paper_25-Enhanced_Resume_Screening_for_Smart_Hiring.pdf?utm_source=chatgpt.com" \o "[PDF] Enhanced Resume Screening for Smart Hiring Using Sentence ...))

**CareerBERT (2025):** Domain-adapted embeddings for resume-to-ESCO job matching via cosine similarity. ([ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0957417425006657?utm_source=chatgpt.com" \o "CareerBERT: Matching resumes to ESCO jobs in a shared ...))

**Resume2Vec (MDPI):** Embedding-driven matching outperforming traditional ATS on human-preference metrics. ([MDPI](https://www.mdpi.com/2079-9292/14/4/794?utm_source=chatgpt.com" \o "Resume2Vec: Transforming Applicant Tracking Systems with ... - MDPI))

**ConFit v2 (2025):** Contrastive/hypothetical finetuning for sparse interaction labels in resume-job datasets. ([arXiv](https://arxiv.org/html/2502.12361v1?utm_source=chatgpt.com" \o "ConFit v2: Improving Resume-Job Matching using Hypothetical ...))

**Hybrid search guidance:** Best-practice patterns for combining lexical + vector signals in retrieval. ([Medium](https://medium.com/google-cloud/hybrid-search-combining-semantic-and-keyword-approaches-for-enhanced-information-retrieval-6a7c046c89ea?utm_source=chatgpt.com" \o "Hybrid Search: Combining Semantic and Keyword Approaches for ...))

**Applicant perceptions & XAI:** Positive applicant attitudes when AI is seen as useful; XAI can change human selection patterns—underscoring need for careful design. ([ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2451958823000362?utm_source=chatgpt.com" \o "Applicants' perception of artificial intelligence in the recruitment ...), [SpringerLink](https://link.springer.com/article/10.1007/s12525-022-00600-9?utm_source=chatgpt.com" \o "Applying XAI to an AI-based system for candidate management to ...))

**Bias & regulation:** EEOC guidance and legal scrutiny of hiring AI tools; history of AI fairness initiatives. ([EEOC](https://www.eeoc.gov/sites/default/files/2024-04/20240429_Employment Discrimination and AI for Workers.pdf?utm_source=chatgpt.com" \o "[PDF] Employment Discrimination and AI for Workers), [Reuters](https://www.reuters.com/legal/transactional/eeoc-says-workday-covered-by-anti-bias-laws-ai-discrimination-case-2024-04-11/?utm_source=chatgpt.com" \o "EEOC says Workday must face claims that AI software is biased))

# ****Research Methodology****

## ****1. Research Design****

This study adopts a **design science research (DSR) approach**, focusing on the iterative development and evaluation of an AI-powered Resume Analyzer. The methodology combines both **qualitative analysis** (understanding recruiter pain points and limitations of existing systems) and **quantitative analysis** (evaluating models on datasets using standard metrics).

The research follows these phases:

**Problem Identification** → Limitations of keyword-based ATS systems.

**Objective Definition** → Develop a hybrid AI-driven system that integrates semantic similarity, keyword analysis, and fairness-aware evaluation.

**Design & Development** → Build prototype using SBERT, keyword matching, and hybrid scoring.

**Demonstration** → Apply system on collected resumes and job descriptions.

**Evaluation** → Use ranking metrics (AUC, Precision@k, nDCG) and ablation studies.

**Conclusion** → Document findings, implications, and future research directions.

## ****2. Data Collection****

### 2.1 Data Sources

**Resumes Dataset:**  
Collected from open-source resume repositories, anonymized resumes from Kaggle datasets, and synthetic resumes generated to reflect diverse skill sets.

**Job Descriptions Dataset:**  
Collected from online job portals (e.g., LinkedIn, Indeed, Glassdoor), focusing on technology and management domains.

### 2.2 Data Preprocessing

**Text Cleaning:** Removal of special characters, formatting inconsistencies, and non-informative sections.

**Tokenization & Normalization:** Lowercasing, stopword removal, lemmatization.

**Skill Extraction:** Identification of skills using predefined taxonomies (e.g., ESCO, O\*NET) and Named Entity Recognition (NER).

**Embedding Generation:** Encoding text using **SBERT embeddings** for semantic similarity.

## ****3. System Architecture & Model Design****

### 3.1 Baseline Methods

**Keyword Matching (TF-IDF):** Traditional ATS-style approach relying on direct keyword overlap.

### 3.2 Proposed Hybrid Model

**Semantic Matching:**

SBERT embeddings for job-resume pairs.

Cosine similarity used to measure semantic closeness.

**Keyword Overlap:**

TF-IDF score for direct keyword matches.

**Hybrid Score:**

Weighted combination of semantic and keyword scores.

Optionally, strictness penalty to reduce inflated keyword matches.

Mathematical Representation:

Score(resume,job)=α⋅Semantic(resume,job)+β⋅Keyword(resume,job)−γ⋅Penalty

## ****4. Evaluation Framework****

### 4.1 Metrics

**AUC (discrimination power)**

**Precision@k (retrieval quality at top-k)**

**nDCG (ranking effectiveness)**

### 4.2 Ablation Study

SBERT-only vs. Keyword-only vs. Hybrid vs. Hybrid+Penalty.

### 4.3 Validation

**Cross-validation** on multiple resume-job datasets.

**Statistical Significance Tests** (e.g., paired t-test) to confirm improvements.

## ****5. Ethical and Fairness Considerations****

Conduct **Fairness Probes** (adverse impact ratio / demographic parity).

Ensure **bias mitigation** by not overfitting to gendered or biased job descriptions.

Use anonymization to protect candidate privacy.

## ****6. Tools and Technologies****

**Programming Language:** Python

**Libraries/Frameworks:**

NLP: Hugging Face Transformers, SpaCy

Similarity: SBERT, cosine similarity

Fairness Auditing: AI Fairness 360, FairLearn

**Dataset Storage & Processing:** Pandas, SQLite/CSV

**Visualization:** Matplotlib, Seaborn

## ****7. Summary****

This methodology enables a structured and replicable framework for developing and evaluating the AI Resume Analyzer. By combining traditional keyword approaches with modern semantic embeddings and fairness considerations, the study aims to design a robust system that can be empirically validated and extended to real-world recruitment scenarios.