

Evaluating Hidden-Content Discrepancies and Ranking Stability in Hybrid AI-Based Resume Screening Systems: A Controlled Experimental Study Proposal

1st Abhishek Ithape

Department of Computer Science
Chandmal Tarachand Bora College
M.Sc. Computer Science

2nd Karan Mhaske

Department of Computer Science
Chandmal Tarachand Bora College
M.Sc. Computer Science

3rd Tanvi Bendre

Department of Computer Science
Chandmal Tarachand Bora College
M.Sc. Computer Science

4th Prof. Amol Sathe

Research Guide
Department of Computer Science
Chandmal Tarachand Bora College

Abstract

Artificial Intelligence (AI) is increasingly used in recruitment to automate resume screening and candidate ranking. These systems process resume documents, extract relevant information, compare candidate profiles with job descriptions, and generate rankings. However, differences in document structure, text extraction, and content representation may affect the information available to a screening system. Such differences may lead to changes in candidate rankings even when the underlying candidate information is intended to remain equivalent. This research proposes a controlled experimental study to evaluate hidden-content discrepancies and ranking stability in hybrid AI-based resume screening systems. The study will compare a baseline screening approach with a hybrid approach combining keyword-based matching and semantic similarity. Controlled resume variations and document-extraction conditions will be investigated to determine whether they influence candidate rankings. Ranking stability will be evaluated using suitable measures, including Spearman's rank correlation, Kendall's rank correlation, top-k overlap, and Normalized Discounted Cumulative Gain (NDCG), where appropriate. The experimental design will include controlled variables, repeated trials where applicable, and statistical analysis of ranking differences. The proposed study aims to develop a reproducible evaluation framework for examining the reliability of automated resume screening. The results, once the experiments have been conducted, may help identify document-processing conditions that require greater attention when designing transparent and dependable AI-assisted recruitment systems. No experimental results are claimed in this proposal. **Index Terms**— Candidate ranking, document

extraction, information retrieval, semantic similarity, stability evaluation.

I Introduction

Artificial Intelligence has emerged as an important technology in modern recruitment. An organization may receive hundreds or even thousands of resumes for a single vacancy. Manually screening every resume requires considerable time and effort. AI-based resume-screening systems can help recruiters extract information from documents and compare candidate profiles with job requirements.

Conventional screening systems often use keyword matching. For example, a system may search for skills such as Python, SQL, Java, or Machine Learning. This approach is relatively simple and easy to interpret, but it may fail to recognize skills expressed using different words or phrases. Other techniques use Natural Language Processing (NLP) and semantic similarity to compare resumes with job descriptions. A hybrid system can combine keyword matching with semantic representations to estimate candidate relevance.

However, resumes are not always presented in a uniform format. PDF files, word-processing documents, columns, tables, and text-extraction software may produce different text orders. Visible information may be extracted incorrectly, omitted, or placed in an unexpected order.

These issues raise an important research question: can discrepancies in resume content extraction influence the ranking of applicants? Ranking stability matters because minor changes in document representation should not cause unjustified changes

precise boundaries of the research contribution.

III Research Problem

Resume-screening systems rely on the information extracted from resume documents. If the same candidate information is extracted differently under different document conditions, the screening software may calculate different relevance scores or produce a different candidate ranking.

The central problem is to determine whether controlled variations in resume information extraction lead to changes in ranking, and whether the size or pattern of those changes differs between a keyword-based method and a hybrid AI-based method. The study will record the original document content, the extracted text, the differences introduced by each condition, the resulting scores, and the candidate rankings.

IV Materials and Methods

A Research Design

An experimental research design will be used. A collection of anonymized or simulated resumes will be compared with selected job descriptions. Candidate profiles will be held constant while document representations and extraction conditions are varied. The same candidate profiles and job descriptions will be used across conditions to support a controlled comparison.

The experiment will include a baseline condition and alternative document-processing conditions. Each condition will be documented so that the experiment can be repeated. Where appropriate, repeated trials will be used to check whether results are consistent.

B Dataset Preparation

The dataset will contain resumes and job descriptions. Candidate profiles may include skills, education, work experience, and other job-related information. If synthetic data are used, profiles will be designed to represent different levels of relevance to the target job descriptions. Criteria for any human relevance judgements will be specified before evaluating the rankings.

Personally identifying information will not be used unless appropriate authorization has been obtained. Dataset preparation will document the source, file format, preprocessing steps, and any changes made to the documents.

C Experimental Conditions

The study will investigate the following conditions:

1. Standard resume text extraction.
2. Resume files represented in different document formats.
3. Controlled changes to text order or document layout.

4. Text-extraction conditions involving omission or incorrect representation of text.
5. Visually hidden or inconsistently displayed text under controlled conditions.

All documents will be subjected to defined conditions. For each case, the original content, extracted text, and differences introduced by the condition will be recorded. The experiment will distinguish intended content from extraction artefacts so that observed ranking changes can be interpreted carefully.

D Baseline Screening Method

The baseline method will rely on keyword matching between the skills in a resume and the requirements in a job description. A simple score can be represented as:

$$S_{\text{keyword}}(r, j) = \frac{|K_r \cap K_j|}{|K_j|}, \quad (1)$$

where K_r is the set of keywords extracted from resume r , and K_j is the set of keywords identified in job description j . The score represents the proportion of job-description keywords that also occur in the resume keyword set. The exact preprocessing steps, keyword lists, and scoring rules will be fixed and documented before the experiment.

E Hybrid AI-Based Screening Method

The hybrid approach will combine keyword matching and semantic matching. The candidate relevance score may be defined as:

$$S_{\text{hybrid}} = \alpha S_{\text{keyword}} + (1 - \alpha) S_{\text{semantic}}, \quad (2)$$

where S_{keyword} is the keyword-based relevance score, S_{semantic} is the semantic similarity score between the resume and job description, and α is a weighting factor between zero and one.

Semantic matching may use sentence embeddings and cosine similarity. The embedding model, weighting factor, preprocessing procedure, and other parameters will be selected based on resource availability and the research objectives. These choices will be recorded to support reproducibility. The study will not assume in advance that the hybrid method is more stable than the baseline method.

F System Architecture

Fig. 1 presents the proposed workflow for evaluating ranking stability.

G Evaluation Metrics

The following metrics will be considered where the available data support their use:

- Spearman's rank correlation: Measures the association between two candidate rankings.

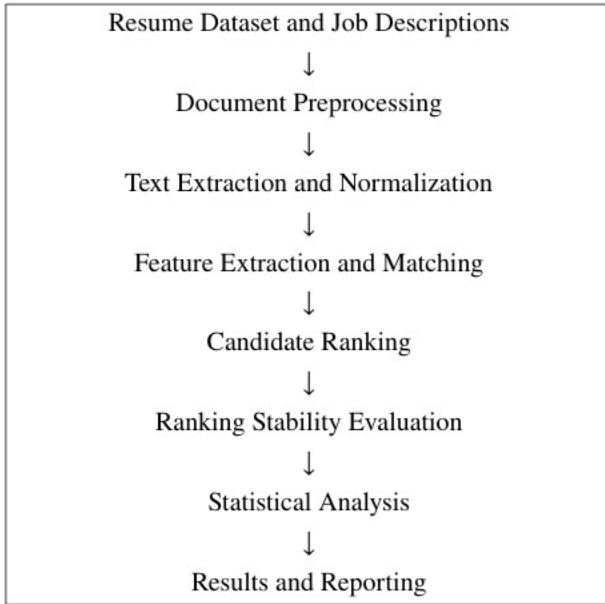


Figure 1: Proposed architecture for resume-ranking stability evaluation

- Kendall's rank correlation: Measures agreement between two rankings using concordant and discordant candidate pairs.
- Top-k overlap: Measures the overlap between the top-ranked candidates under two conditions. It can be written as

$$O_k = \frac{|A_k \cap B_k|}{k}, \quad (3)$$

where A_k and B_k are the sets of top-k candidates from two conditions.

- Normalized Discounted Cumulative Gain (NDCG@k): Evaluates ranked results using relevance ratings and gives greater importance to relevant items appearing near the top.
- Precision@k and Recall@k: These may be calculated when suitable relevance labels are available.

The choice of metrics will be documented along with their interpretation and any assumptions.

H Statistical Analysis

Rankings from paired experimental conditions will be compared. The statistical test will be selected according to the data distribution, dependence structure, and comparison being performed. Descriptive statistics, effect sizes, and confidence intervals will be reported where applicable. Multiple comparisons will be considered when several experimental conditions are evaluated.

No claim of statistical significance or experimental outcome will be made without analysing the collected data. Results will be presented with enough methodological detail to make the analysis understandable and reproducible.

I Tools and Technologies

Possible tools include:

- Python for implementing the experiments.
- pandas and NumPy for data preprocessing and numerical operations.
- scikit-learn for baseline methods and evaluation utilities.
- PDF text-extraction libraries for extracting text from resume documents.
- Sentence-BERT or another documented embedding algorithm for semantic comparison.
- SciPy for statistical analysis.
- Matplotlib for visualizing ranking comparisons.

The selected tools, software versions, parameters, and preprocessing steps will be documented for reproducibility.

J Ethical Considerations

The experiment will follow principles of confidentiality, transparency, and responsible evaluation. Anonymized or simulated resume data will be preferred. Protected characteristics will not be used to discriminate against particular groups. The research will evaluate system behaviour rather than make unsupported claims about individual applicants. AI-generated rankings should not be treated as the sole basis for employment decisions.

V Results and Evaluation Plan

Results will be generated only after the experiments have been performed. Table I provides a template for organizing the planned comparisons; it does not contain experimental findings.

The analysis will compare rankings across conditions and methods. Measured values, confidence intervals, effect sizes, and statistical results will be included after the experiment is completed. The study will examine whether rankings remain stable under the defined conditions and will report cases where the available evidence is inconclusive.

Table 1: Proposed Evaluation Template

Condition	Baseline method	Hybrid approach
Original document	Record ranking and baseline metrics	Record ranking and hybrid metrics
Alternative document form	Measure changes relative to original	Measure changes relative to original
Text-order/layout variability	Compare ranking correlation and top-k overlap	Compare ranking correlation and top-k overlap
Content omission or extraction error	Record changes in extracted keywords and ranking	Record changes in extracted content, similarity, and ranking
Hidden or inconsistently displayed text	Evaluate effects under the defined controlled condition	Evaluate effects under the defined controlled condition

VI Discussion

The proposed study will explore the association between resume-document processing and candidate-ranking stability. If rankings change, the analysis will investigate whether the changes are associated with differences in extracted text, matching scores, or ranking procedures.

Comparing keyword and hybrid approaches will help characterize their behaviour under the same experimental conditions. However, the hybrid approach cannot be considered more reliable in advance; that conclusion must depend on the experimental evidence. Findings may help identify document-validation practices, preprocessing requirements, and systematic tests that are useful when developing AI-supported recruitment systems.

The interpretation will consider the selected dataset, document formats, extraction tools, model parameters, and evaluation metrics. The proposal defines a plan for evaluation and does not report completed experimental results.

VII Limitations and Future Scope of Research

The study may be limited by dataset size, language coverage, document formats, text-extraction tools, and the embedding model used. Synthetic resumes may not fully represent the variety of documents encountered in practical recruitment. Human relevance judgements, if used, may also introduce subjectivity and will require clear criteria.

Future research may examine multilingual resumes, larger and more diverse datasets, alternative ranking models, explainable AI techniques, and human assessment of candidate relevance. Additional work could also investigate how document-processing choices interact with accessibility requirements and responsible recruitment practices. Any conclusions must be interpreted in light of the dataset and experimental design.

VIII Conclusion

This research proposal presents a method for assessing hidden-content discrepancies and ranking stability in hybrid AI-based resume-screening systems. It proposes a comparison between conventional keyword screening and a hybrid method that combines keyword matching with semantic similarity.

Candidate-ranking behaviour will be examined under controlled document-processing conditions using established ranking metrics and statistical analysis. The intended contribution is a reproducible framework for evaluating whether differences in resume content extraction affect candidate ranking. Final conclusions will be drawn only after conducting the experiments and analysing the results.

IX Author Contributions

The proposed contribution statement is provided below and should be reviewed by all authors before final submission.

- Abhishek Ithape: Experimental design and implementation.
- Karan Mhaske: Literature review and dataset preparation.
- Tanvi Bendre: Evaluation and documentation.

The final statement should accurately reflect the actual contributions made by each author.

X Data Availability Statement

The dataset and source code used for the experiments will be described and made available through an appropriate repository, provided that privacy, consent, licensing, and institutional requirements are satisfied. The final statement will identify the repository after it has been selected and the materials are ready for sharing.

Acknowledgment

The authors wish to thank Prof. Amol Sathe, Research Guide, and Chandmal Tarachand Bora College for their support and academic guidance. This acknowledgement should be reviewed and approved by the authors before submission.

References

- [1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008. [Online]. Available: <https://nlp.stanford.edu/IR-book/>
- [2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Language Process. 9th Int. Joint Conf. Natural Language Process. (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.
- [3] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002. doi: 10.1145/582415.582418.