**Original Research Article**

**AN ENHANCED K-NN ALGORITHM LEVERAGING BERT TECHNIQUES FOR RESUME PARSING SYSTEM**

## ABSTRACT

The increasing volume of job applications has created significant challenges for organizations in efficiently screening and ranking candidate resumes. Manual and keyword-based automated systems often struggle with accuracy, and contextual understanding. The study introduced an experimental design that develops a hybrid ensemble model for resume parsing and ranking, combining k-nearest neighbors (KNN) and Bidirectional Encoder Representations from Transformers (BERT). The enhancement lies in BERT's ability to generate deep contextual embeddings that are integrated into KNN’s distance-based classification and keyword matching to improve contextual accuracy, a combination not commonly explored in previous resume parsing systems. The research involved stages such as data cleaning, preprocessing, feature extraction using named entity recognition (NER), model development and training. The system achieved 96.91% parsing accuracy and 100% ranking accuracy across 962 resumes, demonstrating strong performance with precision, recall, and F1-score of 97.0% and allows resumes in DOCX, PDF, or image formats as input. Using Natural Language Processing (NLP) techniques, term frequency- inverse document frequency (TF-IDF) vectorization, and cosine similarity, the system processes resume and ranks them based on relevance to job descriptions with a similarity score. The study was conducted at Air Force Institute of Technology within the time frame of December 2024 and June 2025. The system built highlighted the importance of automated resume parsing systems in recruitment processes.

***Keywords:*** *NLP, KNN, BERT, Vectorization, TF-IDF, Cosine Similarity, Similarity Score, Resumes*

## 1.0 Introduction

The job search process remains a constant reality for many, particularly for new graduates eager to apply their skills and enter the workforce. With diverse job opportunities available today, companies are implementing different strategies for tackling unemployment, leading to the rise of on-site, hybrid, and remote positions (Lamovšek et al., 2024). Every day, companies are looking for many skilled and experienced candidates for different positions regardless of the job type, one common requirement is the submission of resume, which provides a concise representation of a candidate’s qualifications and experience (Joseph et al., 2020). A Resume Parser is a powerful software tool designed to streamline the recruitment process, making it more efficient and easier for both employers and job seekers (Pawar et al., 2024). In recent years Machine Learning (ML) algorithms have been suggested by several researchers to automate the process of resume parsing. Some of the popular techniques used in this regard are those utilizing Natural Language Processing (NLP) techniques (Wahedna et al., 2023). Deep learning models have demonstrated strong potentials due to their ability to identify relevant information from complex, noisy, large and unstructured datasets (Tallapragada et al., 2023). However, existing automated resume parsers, while useful, struggle with format inconsistency, poor contextual understanding, and unstructured data handling, particularly due to their reliance on keyword-based methods. To address these limitations, this research proposes a novel resume parsing system that employs a hybrid ensemble approach combining KNN’s distance-based classification with BERT’s contextual embeddings to provide contextual understanding of resume content when screening and shortlisting candidates based on their similarity scores to the job description. The proposed resume parsing system is highly relevant in today’s competitive job market, where efficient and accurate candidate screening is essential for organizations handling large volumes of applications.

## The rest of the paper is structured as follows: Section 1 presents the Literature Review, which critically examines existing resume parsing techniques and identifies current research gaps. Section 2 describes the Methodology, detailing the development of the KNN-BERT model, including dataset curation, feature extraction, and the overall system architecture. Section 3 covers the Results and Discussion, presenting the model’s classification reports and visual analysis, and benchmarking the proposed system against existing approaches. Section 4 concludes the paper by summarizing key findings, addressing limitations, and suggesting directions for future research.

## 1.1 Literature Review

Bhoir et al. (2023) proposed a hybrid resume parsing system that integrates spacy transformer BERT model and spacy NLP components to extract data and enhance information extraction from unstructured resume. The spacy transformer BERT was used to get the semantic meaning of text and spacy NLP was used for NLP tasks such as tokenization and feature extraction of ten key entities: Name, College Name, Degree, Graduation year, Years of Experience, Designation, Skills, Location Email Adress and Companies worked at. Their methodology began by structuring documents using tools like PDFPlumber and doc2txt, followed by segmenting the data and applying NER. Fuzzy string matching and Regrex pattern matching were employed for similarity checking to check similarity. The authors emphasized the importance of deep learning in this domain and experimented with multiple models such as Conditional Random Fields (CRF), Recurrent Neural Networks (RNN), Convolution Neural Networks (CNN) and Transformer based models. A dataset containing 500 resume were used and reported an accuracy between 90-92% was achieved from the deep learning model. The results indicated that the hybrid approach outperformed existing state-of-the-art resume parsers in terms of both accuracy and efficiency. The study concluded with a recommendation for further research in applying advanced deep learning techniques to improve Human Resource Technologies.

Roy et al. (2020) study was focused on classification and matching of resumes. The system employed Cosine Similarity and KNN for identifying resumes that closely match the input job description, followed by content- based filtering to rank top candidates. The dataset used contained resumes scraped from online job portals and Kaggle organized with three columns: ID, Category and Resume. Preprocessing steps included masking, stemming, lemmatization and stop word removal. The system was tested on different models and among them, the linear SVM performed best achieving an accuracy of 78.52%. The other models used were Random Forest (38.99%), Multinomial Naïve Bayes (44.39%), Logistic Regression (62.40%). Despite the results the model was limited to processing resumes in csv format, but in the real world, the CVs are either in .doc or .pdf formats. Additionally, the use of Genism library led to loss of relevant information due to text summarization, and the system lacked contextual understanding of the resume content.

Fareed et al. (2021) developed a resume classification system using NLP, TF-IDF, KNN and Cosine Similarity. The system used Scikit-learn for ML tasks. Their methodology began with preprocessing steps such as tokenization using NLTK tokenizers, stop word removal, and lemmatization to normalize the data. This was followed by vectorization using TF-IDF, Cosine Similarity was then applied to compute similarity scores between resumes and job description, and KNN was used to classify and rank the resumes. The results achieved were accuracy (98.96%), kappa statistics (98.90%), precision (99.10%), recall (98.90%), F1-score (98.95%). The system’s implementation was limited to resume in CSV format.

Jagwani et al. (2023) integrated Latent Dirichlet Allocation (LDA) for resume ranking and spaCY for entity detection. The system utilized spaCY NER for extraction of details and candidate selection. When focusing on skills the system achieved an accuracy of 77%, but when focusing on all attributes it gave an accuracy of 82%. But there were challenges in entity recognition due to variations in resume formats and heavy reliance on the quality and variety of the dataset used for training the models.

Joseph et al. (2020) proposed a resume screening system that integrates a web-application to assists recruiters in analyzing resumes. The system utilized KNN, Content-Based Filtering to perform fuzzy matching between resumes and job descriptions. The resumes are ranked based on the computed similarity scores and the output is presented as a list of ranked candidates to the recruiter. The methodology began with retrieving the resumes in various formats such as .docx and .pdf, and converting them into plain text using the python library Textract and preprocessing the text with NLTK and Genism Library. The system shows an accuracy of 90.53% for parsing and 77.66 % for scoring. Additionally, it included a skill verification test for candidates and offered course recommendations to those who didn’t meet the specified job criteria, benefitting both recruiters and applicants. The system relied on the test scores which may not actually reflect the candidate’s real-world capabilities.

Tejaswini et al. (2022) used a Multiple-Choice Question (MCQ) test for a particular subject and face detection to prevent malpractice. Those who passed the test got their resume submitted and ranked. Their methodology contained using NLP for text preprocessing, using TF-IDF for vectorization of terms, cosine similarity and KNN for matching candidates resume to job description. A dataset which contained 50 resumes of java developers and project managers was used. The system achieved an average parsing accuracy of 85% and a ranking accuracy of 92%. However, there was potential loss of information during summarization using the genism library, and it relied on generic MCQ assessment and input quality. A small data set which only focused on 2 job roles was used.

Tallapragada et al. (2023) explained that in resume parsing, there is a heavy reliance on word matching techniques and keywords from the words provided in the resume. Hence lacking a contextual understanding of the resume and causing an oversight in the potential candidate for a job. The study utilized BERT for improved resume parsing by extracting contextual meaning from resumes, which are then classified using vectorization and classification algorithms. The results indicated that the BERT-based system provided better accuracy in identifying relevant information in resumes compared to manual methods. However, the limitations include the complexity of the model, dependency on high-quality datasets, and the challenge of handling diverse resume formats.

## 1.2 Research Gap

Despite the significant advancements in resume parsing systems utilizing NLP and ML techniques, several research gaps remain unaddressed. Studies have primarily focused on methodologies such as TF-IDF vectorization, cosine similarity, or KNN for ranking resumes, often resulting in limited contextual understanding and keyword dependency (Tejaswini et al., 2022). Additionally, reliance on small dataset have hindered the effectiveness of these systems. Other approaches heavily depended on keyword matching, which can lead to the oversight of potential candidates simply because they did not use the exact keywords expected by the system and some of the existing approaches were limited to processing resumes in csv format, restricting their ability to handle diverse file formats (e.g., PDF, DOCX). Furthermore, while deep learning models have shown significant promise in improving resume ranking accuracy, their adoption has been constrained by computational complexity and dataset limitations. To overcome these challenges, this study integrates a hybrid approach that utilizes an integration of the KNN Algorithm and the BERT model to address the aforementioned shortcomings in the conventional application of the KNN algorithm to current resume parsing systems. By implementing the contextual understanding ability of BERT, and the pattern recognition ability of KNN, this model aims to enhance parsing accuracy, improve contextual understanding of resumes. This hybrid ensemble approach addresses the limitations identified in previous works, ensuring a more efficient, scalable, and intelligent resume ranking system that can effectively handle diverse resume formats and real-world recruitment challenges.

## 2.0 Methodology

This study adopts an experimental approach to design, develop, and evaluate a hybrid ensemble resume screening and ranking system that integrates the KNN algorithm with BERT-based semantic understanding. The system addresses major shortcomings in existing resume parsing frameworks, including keyword dependency, lack of semantic context.

## 2.1 Data Collection

The dataset used in this study is the Updated Resume Dataset published by Jillani Softech on Kaggle. It consists of 962 resume samples, each labeled with a corresponding job category. The dataset is structured in a CSV format with two main columns: Resume (containing the full candidates CV) and Category (the job title).

## 2.2 Data Preprocessing

To ensure data quality, the data was preprocessed using several NLP steps such as tokenization and NER. The preprocessing steps were utilized for extraction of key entities like skills, experience, and education, while stemming, lemmatization, and normalization cleaned and reduced noise. Afterwards Vectorization of the data was done using TF-IDF to convert resume and job description text into numerical vectors. Finally, data splitting was done in an 80-10-10 ratio for training, testing, and validation, to avoid model bias and overfitting.

## 2.3 Model Development

The proposed system is built on a hybrid ensemble architecture that combines traditional ML with deep learning to enhance resume parsing and candidate ranking. The model integrates three core components:

**KNN Component**: The KNN model ranks resumes by measuring how similar they are to a job description. After resumes are cleaned and standardized through preprocessing steps such as lowercasing, punctuation removal, and stop word filtering, they are transformed into numerical feature vectors TF-IDF. The model classifies resumes by comparing their text representations using TF-IDF and calculates the distance between these representations to categorize them into relevant job fields.

**BERT Component**: enhances the system by understanding the meaning behind resume and job description text. It processes text by breaking it into smaller pieces and analyzing the context of words, capturing relationships like “software engineer” and “developer” even if different terms are used. BERT converts each resume or job description into a compact numerical summary (called an embedding) that represents its overall meaning. This allows the system to handle varied resume formats and messy text, overcoming the limitations of systems that only look for exact keywords. This study utilized the bert-base-uncased variant of the BERT model, which consists of 12 transformer layers, a hidden size of 768, and 12 self-attention heads. The model was loaded using Hugging Face's BertForSequenceClassification class for a sequence classification task.

**Hybrid ensemble Integration**: The model combines BERT and KNN to create an effective resume parsing system. First, BERT converts resumes and job descriptions into numerical embeddings that capture their semantic meaning. KNN then uses these embeddings to measure similarity between each resume and the job description, producing initial predictions. To refine these predictions, the system employs weighted voting, where KNN’s predictions and BERT’s predictions are combined using normalized weights. For each resume, a counter tallies votes, with KNN contributing its weight to its predicted class and BERT contributing its weight to its predicted class; the class with the highest total weight is selected as the final prediction. This weighted ensemble ranks resumes from most to least similar, leveraging BERT’s deep contextual understanding and KNN’s efficient similarity ranking, outperforming keyword-only models by handling diverse resume styles and semantic nuances.

The resume recommendation system follows a structured pipeline designed to process, analyze, and rank resumes based on their relevance to job descriptions. It integrates traditional text vectorization techniques with deep contextual models to enhance classification accuracy. Figure 1 illustrates the end-to-end workflow, from initial resume ingestion and preprocessing to the application of ML models and final result generation. Each stage in the pipeline contributes to refining the recommendation process, ensuring both semantic understanding and keyword relevance are effectively captured.

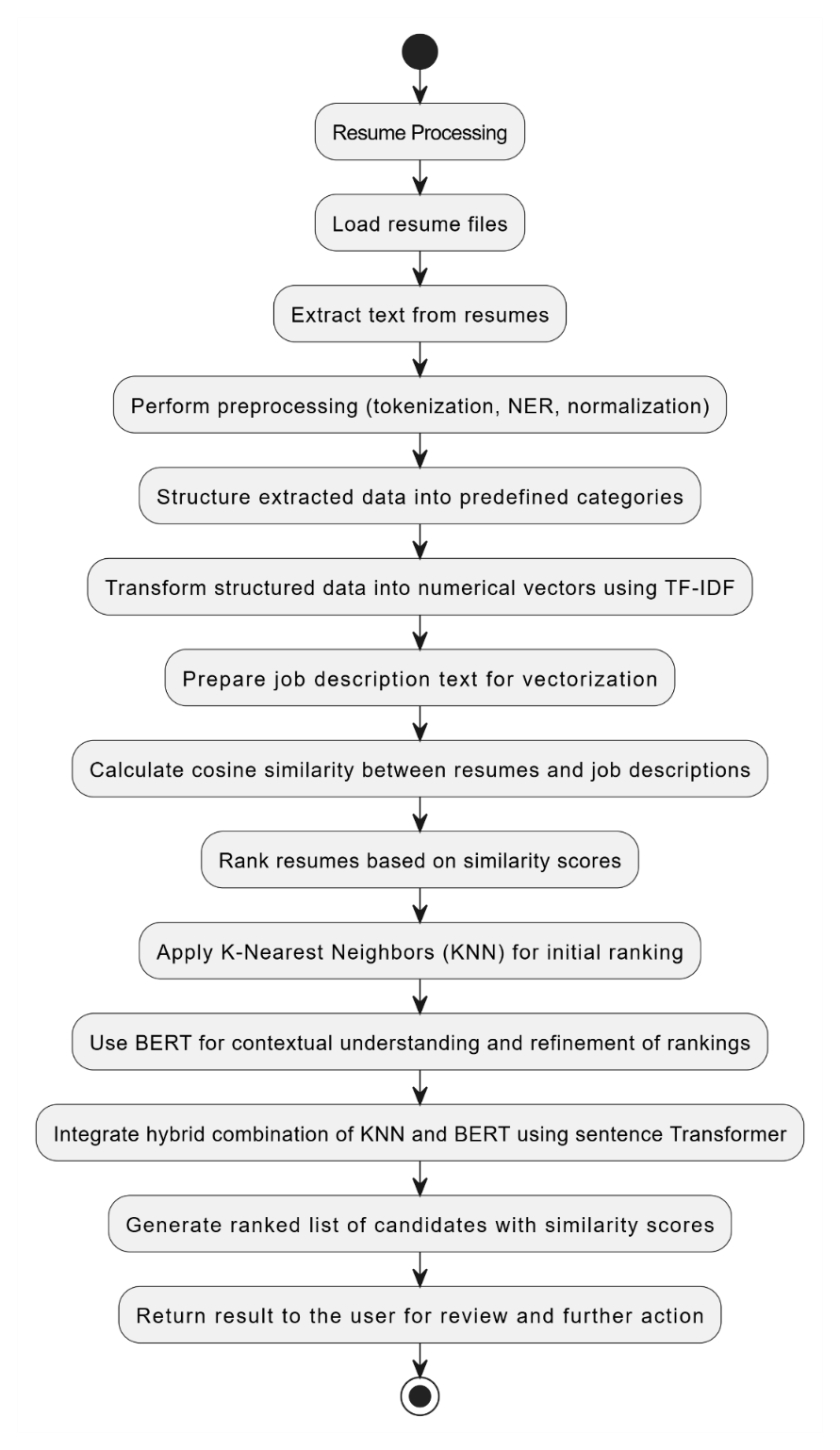


Figure 1: Workflow of the Proposed System

## 3.0 Results and Discussion

This Section presents the comprehensive evaluation results of the system's performance, employing standard metrics such as precision, recall, accuracy, and F1-score. The model was evaluated using Google Colab and the Python programming language. This Section also includes detailed classification reports for the KNN, BERT, and hybrid ensemble models, along with their respective confusion matrices. Furthermore, it benchmarks the system's performance against existing approaches and provides a visual representation of the overall system performance.

## 3.1 KNN Model Evaluation

The model operated in TF-IDF vector space and used cosine similarity to compute resume-job description closeness. With K=11 (determined via accuracy tuning), to further assess the classification performance of the model, a classification report and confusion matrix were generated to indicate how accurately the model predicted each class and where misclassifications occurred, the report shows that the model achieved 92.78% test accuracy as seen in Figure 2 and Figure 3.

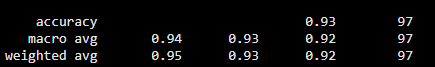


Figure 2: Classification Report of KNN

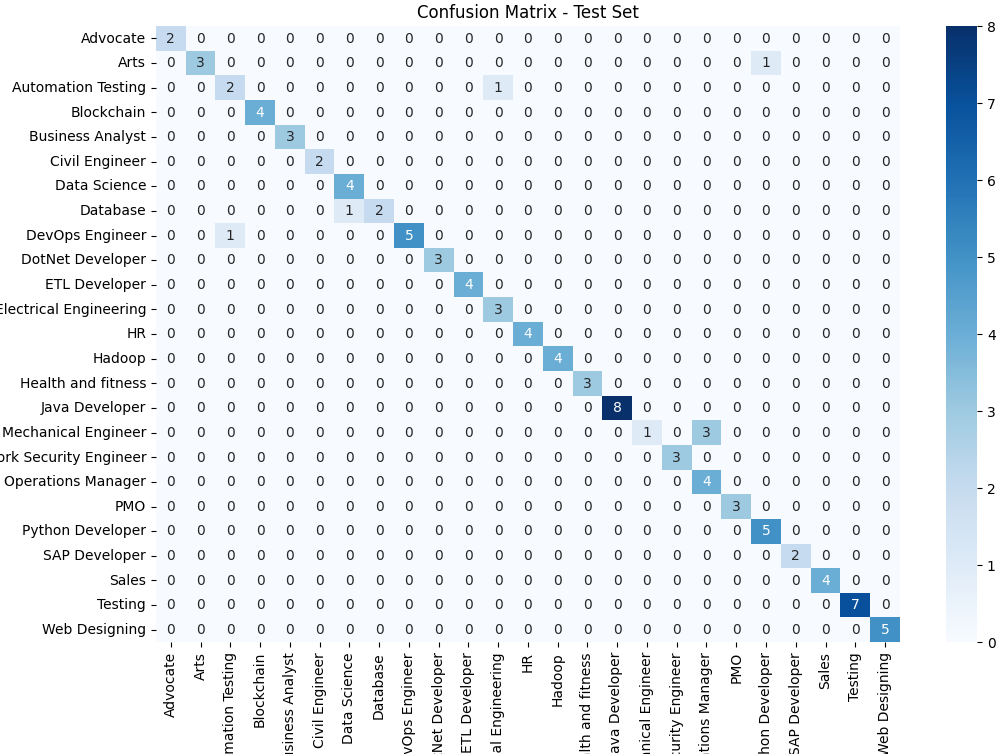


Figure 3: Confusion Matrix of KNN

## **3.2 Evaluation of BERT Model**

Leveraging BERT-based uncased version of pre-trained contextual embeddings, BERT improved semantic understanding and ranking accuracy beyond keyword matching. The BERT model outperformed KNN due to its ability to understand the contextual meaning of words within a sentence, rather than relying solely on surface-level keyword matching. While KNN depends on vector similarity using simple techniques such as TF-IDF or cosine distance, BERT uses deep contextual embeddings derived from a pre-trained transformer network. Its evaluation shows how it outperformed KNN in capturing nuanced resume information giving an accuracy of 97.0% as seen in Figure 4.

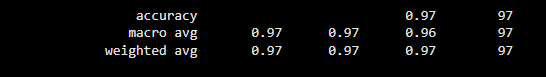


Figure 4: Classification Report of BERT

3.3 Hybrid Ensemble Model

To refine these predictions, the system applies weighted voting, combining KNN’s predictions (weighted by its 0.92 accuracy) and BERT’s predictions (weighted by its 0.97 accuracy) using normalized weights (KNN: 0.92 / (0.92 + 0.97), BERT: 0.97 / (0.92 + 0.97)). A counter sums the weighted votes for each class, selecting the highest-weighted class as the final prediction. It captured both surface-level keywords and deep context for highly accurate ranking giving an accuracy of 97.0%.

The classification report and confusion matrix as seen in Figures 5 and 6 show the performance of the model.

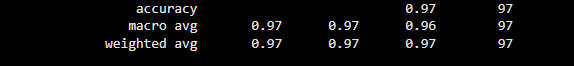


Figure 5: Classification Report of Hybrid Ensemble Model

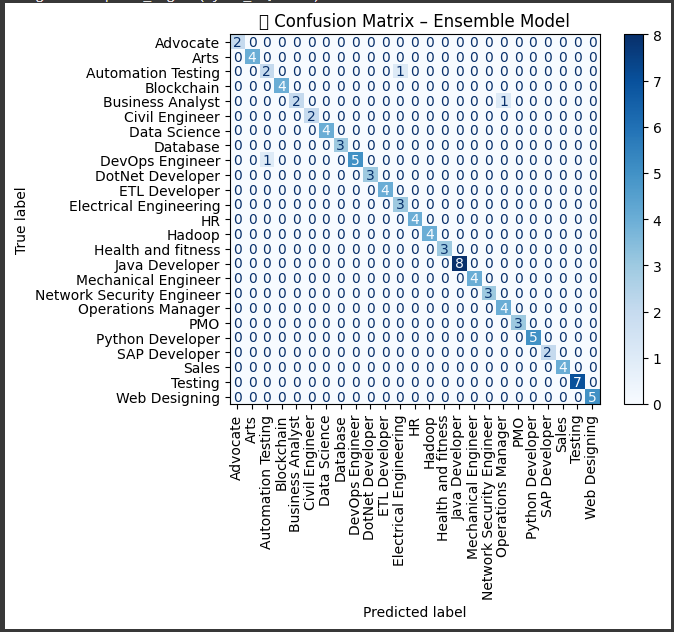


Figure 6: Confusion Matrix of Hybrid Ensemble Model

## 3.4 Algorithm Performance Metrics

The system's effectiveness was evaluated using 962 resumes, focusing on both parsing and ranking performance. For parsing, it achieved an average precision, recall, and F1 score of 97%, with an overall parsing accuracy of 96.91%, indicating high accuracy in identifying key resume elements. For resume ranking, the system reached a ranking accuracy of 100.0% as shown in Figure 7.

Table 1 shows the entire systems performance when evaluated using Parsing Accuracy, Ranking Accuracy, Precision, F1-score and Recall.

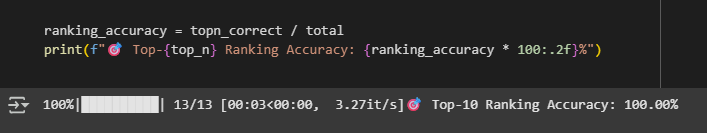


Figure 7: Ranking Accuracy Result

Table 1: Evaluation Results

|  |  |
| --- | --- |
| Total | 962 |
| Parsing Accuracy | 96.91% |
| Ranking Accuracy | 100.00% |
| Precision | 97.00% |
| Recall | 97.00% |
| F1- Score | 97.00% |

## 3.5 Benchmark Comparison

To evaluate the performance of the proposed resume recommendation system, a benchmark comparison was carried out against several existing models from recent literature. Table 2 presents a detailed overview of algorithms employed by other authors, such as TF-IDF with Logistic Regression, Naïve Bayes, KNN, BERT-based models, and content-based filtering techniques. These models varied in terms of evaluation metrics, some emphasized parsing accuracy, while others assessed classification or ranking performance.

Table 2: Benchmarking the System with Existing Systems

|  |  |  |
| --- | --- | --- |
| **Author(s)** | **Algorithms** | **Performance evaluation** |
| Mittal et al. (2020) | TF-IDF  Logistic Regression | Accuracy-91.47%  Precision- 82.0%  Recall- 77.2%  F1 Score- 79.0% |
| Onukwugha et al. (2024) | NLP  Multinomidal Naïve Bayes  KNN | Accuracy-87.0% |
| Joseph et al. (2020) | KNN  Content Based Filtering | Parsing Accuracy-90.53%  Scoring Accuracy-77.6% |
| Nadia (2023) | BERT based NER  VADER Sentiment Analysis  XGBOOST  TOPSIS | XGB Classifier Accuracy on training set-98.0%  XGB Classifier on test set-87% |
| Jagwani et al. (2023) | Latent Dirichlet Allocation  spaCY NER | Accuracy(skills)- 77%  Accuracy(All attributes)-82.0% |
| Tejaswini et al. (2022) | KNN  TF-IDF  Cosine Similarity | Parsing Accuracy- 85.0%  Ranking Accuracy-92.0% |
| Proposed System | KNN  BERT  TF-IDF  Cosine Similarity | Parsing Accuracy- 96.91%  Ranking Accuracy- 100.00%  Precision- 97.00%  Recall – 97.00%  F1-Score- 97.00% |

## 4. Conclusion

This study aimed to develop a resume parsing and ranking system using a hybrid ensemble of KNN-BERT model. By combining KNN similarity with BERT's contextual embeddings, the system was able to accurately extract, classify, and rank resumes according to job relevance. The system achieved a parsing accuracy of 96.91% and a ranking accuracy of 100.00%. The integration of ML and deep learning techniques in an ensemble model resulted in reliable and interpretable predictions.

However, BERT’s computational requirements significantly increased training times and restricted the model to 6 epochs, potentially capping performance improvements. The study concludes that combining complementary algorithms like KNN and BERT significantly enhances performance in complex NLP tasks such as resume parsing. It underscores the importance of hybrid approaches in achieving both accuracy and robustness in recruitment tools. For future improvements, expanding the dataset across various industries, optimizing KNN’s k-value for specific datasets, exploring alternative lightweight classifiers (e.g., SVM) with BERT embeddings and incorporating recruiter feedback can further optimize system performance.

## DISCLAIMER (ARTIFICIAL INTELLIGENCE)

The Authors of this study hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

**References**

Bhoir, N., Jakate, M., Lavangare, S., Das, A., & Kolhe, S. (2023). *Resume parser using hybrid approach to enhance the efficiency of automated recruitment processes.* Datta Meghe College of Engineering. DOI:10.22541/au.168170278.82268853/v1

Deshmukh, A., & Raut, A. (2024). Applying BERT-based NLP for automated resume screening and candidate ranking. *Annals of Data Science*. [DOI: 10.1007/s40745-024-00524-5](https://doi.org/10.1007/s40745-024-00524-5)

Fareed, R. T., Rajath, V., & Kaganurmath, S. (2021). Resume classification and ranking using KNN and cosine similarity. *International Journal of Engineering Research & Technology (IJERT)*, *10*(8).

Jagwani, V., Meghani, S., Pai, K., & Dhage, S. (2023). Resume evaluation through latent Dirichlet allocation and natural language processing for effective candidate selection. *arXiv*. DOI:10.48550/arXiv.2307.15752

Joseph, J., Sunny, J., R, R., Byju, B. E., & K C, L. (2020). Resume analyser: Automated resume ranking software*. International Journal for Research in Applied Science and Engineering Technology* 8(7):896 DOI:10.22214/ijraset.2020.30378

Lamovšek, A., Radević, I., Mohammed, S. S., & Černe, M. (2024). Beyond the office walls: Work design configurations for task performance across on‐site, hybrid and remote forms of work*. Information Systems Journal.* DOI: 10.1111/isj.12542

Li, K., Zhu, M., Zhang, X., & Xia, H. (2024). A gradient-enhanced decision tree and XGBoost based human-job matching model. *Proceedings of the 2024 IEEE Conference on Artificial Intelligence and Networking in Industry and Technology (AINIT)*. DOI:10.1109/AINIT61980.2024.10581848

Mgarbi, H., Chkouri, M. Y., & Tahiri, A. (2023). Towards a new job offers recommendation system based on the candidate resume. *International Journal of Computing and Digital Systems, 14*(1), Article 140103. DOI:10.12785/ijcds/140103

Nisha, B., Manobharathi, V., Jeyarajanandhini, B., & Sivakamasundari, G. (2024). HR Tech Analyst: Automated resume parsing and ranking system through natural language processing. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. DOI:10.1109/ICACRS58579.2023.10404426

Onukwugha, C. G., Ofoegbu, C. I., Aliche, O. B., & Betrand, C. U. (2024). Resume optimization model using machine learning techniques. *International Journal of*

*Intelligent Information Systems, 13*(5), 109-116. DOI:10.11648/j.ijiis.20241305.12

Pawar, A., Kosabe, S., Warde, A., & Mhamunkar, K. (2024). Resume parser. *International Journal of Advances in Engineering and Management (IJAEM), 6(4),* 1147-1153. DOI:10.35629/5252-060411471153

Satheesh, K., Jahnavi, A., Iswarya, L., Ayesha, K., Bhanusekhar, G., & Hanisha, K. (2020). Resume ranking based on job description using SpaCy NER model*.* *International Research Journal of Engineering and Technology*, 7(5), 74–77

Tallapragada, V. V. S., Raj, V. S., Deepak, U., Sai, P. D., & Mallikarjuna, T. (2023). Improved resume parsing based on contextual meaning extraction using BERT. In *Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1702-1708). IEEE. DOI: 10.1109/ICICCS56967.2023.10142800

Tejaswini, K., Umadevi, V., Kadiwal, S. M., & Revanna, S. (2022). Design and development of machine learning based resume ranking system. 371–375. DOI:10.1016/j.gltp.2021.10.002

Tülümen, N., Akgün, G., Nohutçu, A., Aktoros Genç, G. S., & Genç, S. (2021). Hybrid job and resume matcher. *Proceedings of the 2021 6th International Conference on Computer Science and Engineering (UBMK)*, 163–168.

DOI:10.1109/UBMK52708.2021.9558932

Ransing, R., Mohan, A., Emberi, N. B., & Mahavarkar, K. (2021). Screening and ranking resumes using stacked model*.* *In* *Proceedings of the 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)* (pp. 643–648). IEEE.

DOI:10.1109/ICEECCOT52851.2021.9707977

Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020**).** A machine learning approach for automation of resume recommendation system. *Procedia Computer Science, 167*, 2318–2327. DOI:10.1016/j.procs.2020.03.284

Wahedna, A., Vakil, A., Shah, S., Kelkar, V. V., & Shrivastava, I. (2023). *Resume screening–Testing for data stability*. In *Proceedings of the 2023 International Conference on Software Engineering and Systems (ICSES)* (pp. 1–8). IEEE. DOI:10.1109/ICSES60034.2023.10465529

Ransing, R., Mohan, A., Emberi, N. B., & Mahavarkar, K. (2021). Screening and ranking resumes using stacked model*.* *In* *Proceedings of the 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)* (pp. 643–648). IEEE.

DOI:10.1109/ICEECCOT52851.2021.9707977