interAvailable online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 6 (2025) 1329 - 1338 e-ISSN: 2580-0760

Benchmarking Machine Learning Paradigms for Resume Screening on Imbalanced Data

Fitri N. Febriana[1,3], Ira Puspitasari[2,3]
[1]Master Program of Human Resource Development, Faculty of Graduate School, Universitas Airlangga, Surabaya, Indonesia
[2]Information System Study Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[3]Research Center for Quantum Engineering Design, Faculty of Science and Technology, Universitas Airlangga, Surabaya

[1]fitri.noor.febriana-2024@pasca.unair.ac.id, [2]ira-p@fst.unair.ac.id

Abstract

Manual resume screening is an inefficient and bias-prone process, yet comprehensive benchmarks of machine learning models on imbalanced, real-world recruitment data remain scarce. This study addresses this gap by benchmarking seven models from classical, ensemble, and deep learning paradigms for automated resume classification. Using a private dataset of 2,483 resumes across 24 job categories, this study evaluates the models with distinct TF-IDF and BERT embedding feature pipelines and an adaptive strategy for handling class imbalance (Class Weights, SMOTE, SMOTEENN). The results showed that the XGBoost model achieved the highest performance (weighted F1-score of 0.779), followed by the highly competitive BERT (F1 0.728) and Random Forest (F1 0.711) models. Despite these methods, all models struggled with extreme minority classes, confirming data scarcity as a primary limitation. This study provides a valuable benchmark and an evidence-based framework for HR practitioners, highlighting the critical trade-off between predictive performance (XGBoost), interpretability (Random Forest), and semantic capability (BERT). The findings conclude that the primary challenge is data representation, steering future work towards data augmentation and fairness audits.

Keywords: ensemble methods; human resource analytics; imbalanced data; machine learning benchmark; resume classification

How to Cite: F. N. Febriana, and I. Puspitasari, "Benchmarking Machine Learning Paradigms for Resume Screening on Imbalanced Data", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 6, pp. 1329 - 1338, Dec. 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i6.7123

Received: August 25, 2025
Accepted: November 16, 2025
Available Online: December 7, 2025

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The exponential growth of job applicants in the digital era has intensified the need for efficient resume screening systems in human resource (HR) departments. Traditional manual screening processes are time-consuming, prone to human bias, and inefficient for large-scale recruitment [1] - [3]. To address this, organizations increasingly adopt artificial intelligence (AI)-driven tools, ranging from classical machine learning (ML) algorithms, advanced ensemble methods, to Transformer-based deep learning models such as BERT to automate resume classification [4], [5]

While these models are promising, comparative evaluations that measure their performance, efficiency, and interpretability in the context of real-world, imbalanced recruitment data remain limited [6]. A primary challenge is the trade-off between the predictive accuracy of advanced "black-box" models and the practical needs of HR workflows [7] - [9], such as computational efficiency and interpretability [10]. Although traditional models like Random Forest (RF) are known for their transparency [11], [12], they are rarely benchmarked directly against more complex counterparts like gradient boosting or transformers in resume screening tasks [13] - [15]. This results in a lack of comprehensive, comparative evidence to guide the selection of appropriate AI tools in HR

This gap is concerning because, without a comprehensive benchmark, HR departments risk making uninformed decisions when adopting AI technologies [16] - [19]. They may invest in computationally expensive and uninterpretable deep learning models that are ill-suited to their operational scale and fail to meet fairness and transparency standards [20]. Conversely, they might rely on overly simplistic models that lack the necessary predictive power, thus failing to identify the best candidates [21]. This lack of evidence-based guidance hinders the

effective and responsible adoption of AI in recruitment, potentially leading to inefficient resource allocation and perpetuating biases, which undermines the goal of equitable hiring [22].

This study aims to fill this gap by conducting a systematic benchmark of seven ML models, spanning the classical (Naive Bayes, SVM, Logistic Regression), ensemble (Random Forest, XGBoost), and transformer (BERT) categories, using a curated resume dataset. The study utilizes a dataset comprising 2,483 resumes annotated into 24 job categories, which exhibits significant class imbalance. By applying two distinct feature extraction workflows [23], [24] , TF-IDF for the classical and ensemble models [25], [26], and contextual embeddings for the transformer model [27], this research evaluates each model based on its performance, scalability, and potential interpretability [28]. This analysis allows for a fair comparison between the ability of statistical-based models to capture key patterns and the capacity of semantic models to understand contextual nuances.

Through this evaluation, this research provides several contributions. Empirically, the study presents a comprehensive performance benchmark that can serve as a reference for future research in the field of HR analytics. Practically, our findings offer an evidence-based framework to assist HR professionals in selecting context-appropriate models, balancing predictive accuracy with operational constraints such as interpretability requirements and computational resources. As such, this research is relevant to the broader effort to support the goal of SDG 8 in promoting productive employment opportunities for all.

2. Methods

This methodology section outlines a systematic approach to evaluating the performance of seven machine learning models in resume classification. This study adopts the Cross Industry Standard Process for Data Mining (CRISP-DM) [29], [30] to systematically evaluate and compare models spanning classical (Naive Bayes, SVM, Logistic Regression), ensemble (Random Forest, XGBoost), and deep learning (Neural Networks, BERT) paradigms. As illustrated in Figure 1, the methodology follows six interconnected phases: (1) Business Understanding, (2) Data Understanding, (3) Data Cleansing and Preprocessing, (4) Modeling, (5) Evaluation, and (6) Deployment.

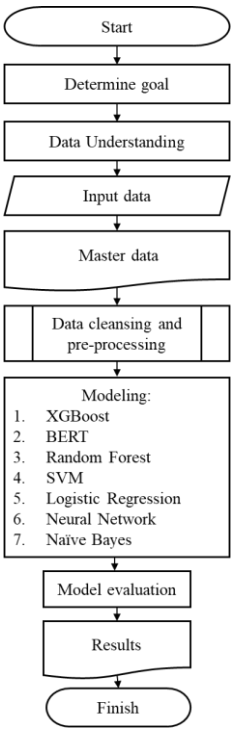


Figure 1. CRISP-DM methodology implementation.

2.1 Data Gathering dan Understanding

The dataset used in this research consists of 2,483 resumes annotated into 24 distinct job categories, sourced from PDF files on public recruitment platforms. Each resume was manually validated by HR professionals to ensure labeling accuracy. An analysis of the class distribution, as shown in Figure 2.

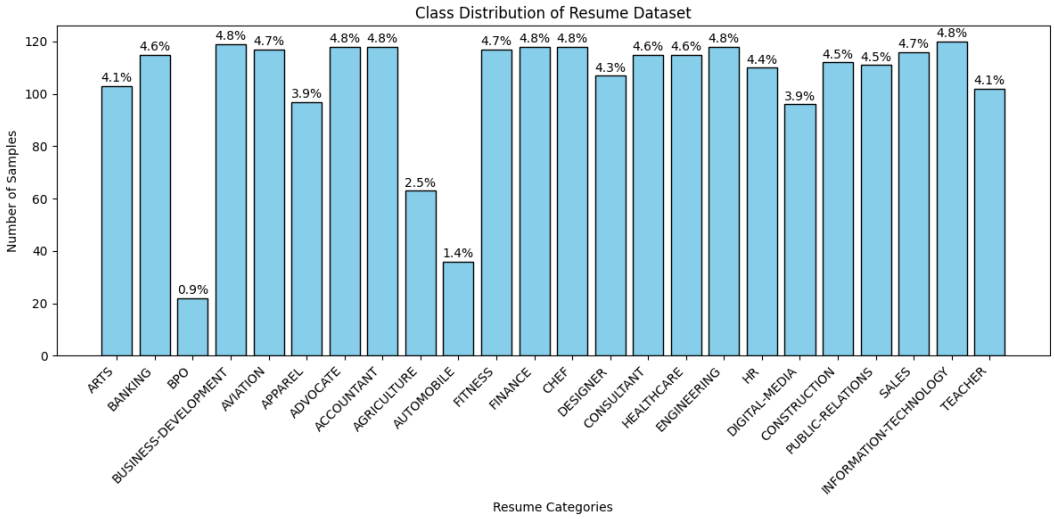


Figure 2. Class distribution of the resume dataset used in the research, showing the number of samples for each of the 24 job categories.

Figure 2, revealed a significant imbalance among categories, with BPO (0.89%, 22 samples) and AUTOMOBILE (1.45%, 36 samples) being the most underrepresented classes, while INFORMATION-TECHNOLOGY (4.83%, 120 samples) was the largest category. The specific names of the recruitment platforms are intentionally withheld to protect data confidentiality. Given the sensitive nature of the data containing personal information, this dataset cannot be made publicly available to protect candidate privacy.

2.2 Data Cleansing and Preprocessing

This section details the comprehensive data preparation pipeline designed to transform raw resume text into a format suitable for the various machine learning models.

The workflow is divided into three primary stages: (1) Text Preprocessing to systematically clean and normalize the textual content; (2) Feature Engineering and Data Splitting to convert text into numerical features and partition the data; and (3) Handling Class Imbalance to mitigate the effects of the skewed class distribution present in the dataset.

2.2.1 Text Preprocessing

The text preprocessing stage, as depicted in Figure 3, is designed to systematically clean and normalize the raw text extracted from PDF resumes to ensure data integrity.

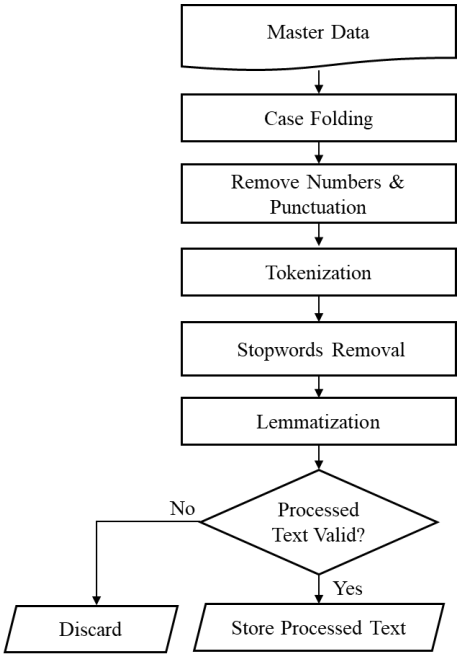


Figure 3. Text preprocessing workflow diagram.

The workflow commences with an initial Text Validation step to identify and filter out any empty or invalid raw text inputs; entries that fail this check are immediately discarded. Valid text then proceeds through a sequential pipeline of NLP transformations. This pipeline includes: (1). Case Folding to ensure case uniformity, (2). Removal of Numbers and Punctuation to eliminate non-textual characters, (3). Tokenization to segment the text into discrete word units, (4). Stopwords Removal to filter out non-discriminative common terms, (5). Lemmatization to reduce words to their fundamental root forms [31].

Following these transformations, a final validation check is performed to confirm that the processed text is still valid and not empty. Any text that becomes invalid during processing is also discarded. Only the successfully processed and validated texts are stored, and their corresponding labels are updated for the subsequent feature engineering stage. This robust workflow ensured high data quality, ultimately retaining all 2,483 resumes as valid samples.

2.2.2 Feature Engineering and Data Splitting

To convert the processed text into a numerical format, the TF-IDF vectorization method was applied [32], as shown in Figure 4.

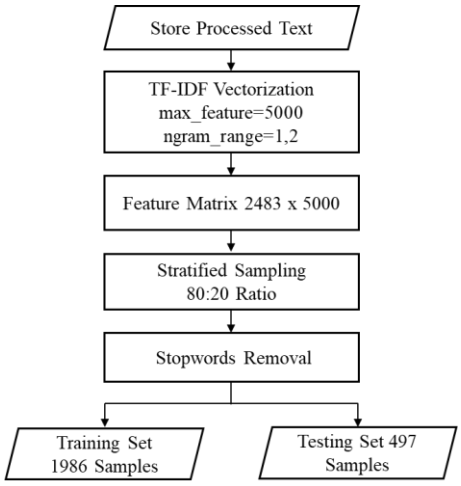


Figure 4. Feature engineering and data splitting workflow diagram.

The configuration was set to max_features=5,000 and ngram_range=(1,2) to capture both single keywords (unigrams) and key phrases (bigrams), such as "project_management". This process resulted in a feature matrix of dimensions (2483, 5000). This dimensionality signifies that each of the 2,483 resumes in the dataset is represented by a vector of 5,000 numerical features, where each feature corresponds to the TF-IDF score of a specific unigram or bigram.

Subsequently, the dataset was split into training and testing sets at an 80:20 ratio, yielding 1,986 samples for training and 497 for testing. This division was performed using stratified sampling to ensure the imbalanced class distribution was preserved proportionally in both sets.

2.2.3 Handling Class Imbalance

To address the significant class imbalance identified in the dataset (Figure 2), a conditional, multi-technique strategy was implemented, as illustrated in Figure 5. This approach applies a specific handling method based on a calculated imbalance ratio: a class weighting strategy was used for low imbalance (ratio < 1.5), the SMOTE (Synthetic Minority Over-sampling

Technique) was applied for moderate imbalance ($1.5 \leq$ ratio < 3) [33], and the SMOTEENN hybrid technique was designated for severe imbalance (ratio ≥ 3)) [34].

Furthermore, to handle potential computational issues arising from the sparse nature of the TF-IDF data, a fallback mechanism was implemented. If a complex resampling technique (SMOTE or SMOTEENN) failed during the fitting process, the system would automatically revert to the more stable class weights method to ensure the training pipeline's continuity. This adaptive and robust strategy ensures that the most appropriate and stable technique is applied based on both the severity of the imbalance and the computational constraints.

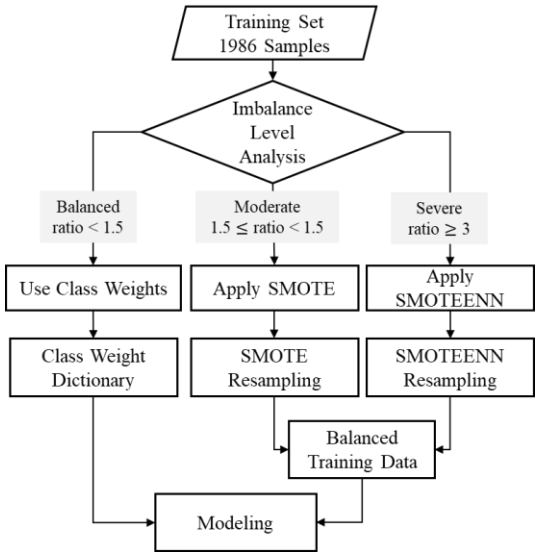


Figure 5. Handling class imbalance workflow diagram.

2.3 Modeling

The modeling phase involved training and evaluating seven models spanning three distinct algorithmic paradigms to ensure a comprehensive evaluation. The classical approaches included Support Vector Machine (SVM) [35], Naive Bayes, and Logistic Regression. Ensemble methods comprised Random Forest [36] and XGBoost [37], while deep learning approaches utilized a Multi-Layer Perceptron (MLP) Neural Network [38], and a pre-trained BERT model [39], [40]. All models were trained on the dataset split into 1,986 training and 497 testing samples. To address the significant class imbalance in the training set, the adaptive strategy detailed in Section 2.2.3 was applied.

Hyperparameters for each model were optimized using a grid search with 5-fold cross-validation on the training set [41] - [43]. The models were implemented using Scikit-learn, TensorFlow (for the MLP), and the Hugging Face Transformers library (for BERT). A consistent random seed was used throughout all experiments to ensure reproducibility.

2.4 Model Evaluation

The performance of the seven trained models was quantitatively assessed on the unseen test set, which consists of 497 samples. This evaluation provides an objective measure of each model's ability to generalize new data. The standard metrics for multi-class classification were used: Accuracy, Precision, Recall, and F1-Score [44] - [46].

To account for the significant class imbalance in the dataset, Precision, Recall, and F1-Score were computed using a weighted average. This method calculates the metrics for each class and finds their average, weighted by the number of true instances for each class (the support). This approach provides a more robust and fair assessment of performance than a simple macro-average, as it appropriately considers the contribution of each class.

2.5 Deployment Plan

The deployment plan proposes a phased approach tailored to an organization's scale and requirements [47]. For Small and Medium-sized Enterprises (SMEs) that prioritize transparency, the Random Forest model is recommended for its high interpretability and can be deployed via a REST API for straightforward integration with Applicant Tracking Systems (ATS).

Conversely, for large-scale enterprises targeting maximum predictive performance, the XGBoost model is the primary choice, to be deployed on a scalable cloud infrastructure and supported by MLOps practices for continuous monitoring [48]. Regardless of the selected model, a responsible deployment must include rigorous bias audits [49], explainability tools such as SHAP/LIME [50], and operate under a human-in-the-loop principle, where the system functions as a decision-support tool rather than the final decision-maker [51].

3. Results and Discussions

This section presents the key findings from the resume classification experiment, focusing on a comparative analysis of seven machine learning models across three different paradigms. The analysis covers: (1) a comparison of overall model performance, concisely visualized in a table; (2) an in-depth, class-level analysis to understand the strengths and weaknesses of each model; and (3) a discussion of the practical implications of these findings.

3.1 Comparative Model Performance

The quantitative evaluation of the seven models on the test set reveals a clear performance hierarchy, as visualized in Tables 1 and 2. The XGBoost model consistently delivered the most superior performance across all measured metrics, achieving an accuracy of 78.9%, a weighted precision of 78.4%, a weighted recall of 78.9%, and a weighted F1-score of 0.779. This dominance indicates that for this dataset, the gradient boosting algorithm was most effective at modeling the relationship between TF-IDF features and job categories. This finding is consistent with other recent studies, which also demonstrate the superiority of

ensemble models like XGBoost for handling imbalanced classification tasks due to their robustness and efficiency [52].

Table 1. Research model nomenclature.

Model	Identity
XGBoost	1
BERT	2
Random Forest	3
SVM	4
Logistic Regression	5
Neural Network	6
Naive Bayes	7

Table 2. Quantitative Evaluation of Model Performance

Model Id.	Accuracy	Precision	Recall	F1-Score
1	0.789	0.784	0.789	0.779
2	0.734	0.736	0.734	0.728
3	0.734	0.735	0.734	0.711
4	0.700	0.700	0.700	0.692
5	0.700	0.710	0.700	0.691
6	0.652	0.648	0.652	0.637
7	0.612	0.631	0.612	0.591

In the second tier, there was a close competition between BERT and Random Forest. Both achieved an identical accuracy of 73.4%. However, a deeper analysis of the F1-score metric shows that BERT (0.728) slightly outperformed Random Forest (0.711). This slight advantage is largely attributable to BERT's marginally higher precision (73.6%) compared to Random Forest (73.5%), while both had the same recall value (73.4%). This suggests that a positive prediction from BERT is slightly more likely to be correct, reflecting a better generalization capability from the transformer model. This finding is consistent with research in the recruitment domain, which indicates that deep learning models leveraging contextual embeddings tend to outperform traditional algorithms due to their superior ability in capturing semantic nuances [53]. Further analysis of the middle tier also reveals interesting nuances. SVM and Logistic Regression both recorded an accuracy of 70.0% and an F1-score of 0.691. However, Logistic Regression exhibited a noticeably higher precision (71.0%) than SVM (70.0%), implying that this model is more conservative and generates fewer false positives. In the lower tier, Naive Bayes demonstrated a unique pattern: its precision (63.1%) was higher than both its accuracy and recall (both 61.2%). This indicates that while Naive Bayes failed to identify many relevant candidates (low recall), the predictions it did make tended to be more reliable (higher precision). Meanwhile, the Neural Network (MLP) (F1-score of 0.637) showed a balanced performance across all metrics without a strong bias towards either precision or recall. These findings underscore the importance of looking beyond the F1-score to understand the specific behaviors and trade-offs of each model.

3.2 Class-Level Performance Analysis

A more in-depth analysis of each model's classification report reveals consistent patterns related to the handling of different job categories, particularly the contrast between majority and minority classes. The complete classification reports for all seven models, which provide a detailed breakdown of per-class metrics, are available in the Supplementary Information (see Table S1-S7). These findings highlight data scarcity as a primary limiting factor and demonstrate that the top-performing models possess distinct strengths.

Nearly all models exhibited significant difficulty in classifying categories with a very small number of samples. The BPO category, with only four samples in the test set, consistently yielded an F1-score of 0.00 for almost all models, including XGBoost (Table S1) and Random Forest (Table S3). The sole exception was BERT (Table S2), which managed to achieve an F1-score of 0.22. This suggests that the pre-trained knowledge within BERT provides it with some resilience even on extremely scarce data, where other models fail completely. Poor performance was also observed for the AUTOMOBILE (7 samples) and AGRICULTURE (13 samples) categories, which consistently received low F1-scores across all models. This confirms that without a sufficient volume of data, models struggle to learn representative patterns, regardless of the imbalance-handling techniques applied.

On the other hand, the high-performing models demonstrated unique strengths in categories with adequate data. XGBoost (Table S1) showed its dominance in categories with distinct and technical vocabularies, achieving outstanding F1-scores in classes like CONSTRUCTION (0.94), AVIATION (0.92), and ACCOUNTANT (0.92). This superiority demonstrates XGBoost's proficiency in maximizing the predictive signal from the keyword-based features extracted by TF-IDF. In contrast, BERT (Table S2) excelled in categories requiring nuanced semantic understanding and contextual overlap. Its exceptionally high performance in categories such as BUSINESS-DEVELOPMENT (0.94) and CONSULTANT (0.89) proves its ability to distinguish between roles with often similar job descriptions. This capability stems from its transformer architecture, which can understand the meaning of words based on sentence context—a significant advantage over models that rely solely on word frequency.

Further analysis also uncovers interesting behaviors in other models. For instance, on the ACCOUNTANT category, Random Forest (Table S3) exhibited a perfect recall (1.00) but a lower precision (0.65). In practical terms, this means Random Forest successfully identified all of the actual accountant candidates but also incorrectly labeled many candidates from other classes as accountants. This highlights a classic precision-recall trade-off: the model can serve as a good "safety net" for not missing potential candidates, but at the cost of requiring more subsequent manual filtering. This finding underscores that model selection depends

not only on the overall F1-score but also on the specific business objectives within the recruitment process.

To visualize these misclassification patterns in greater depth, the confusion matrices of the two best-performing models (XGBoost and BERT) are presented in Figure 6 and Figure 7.

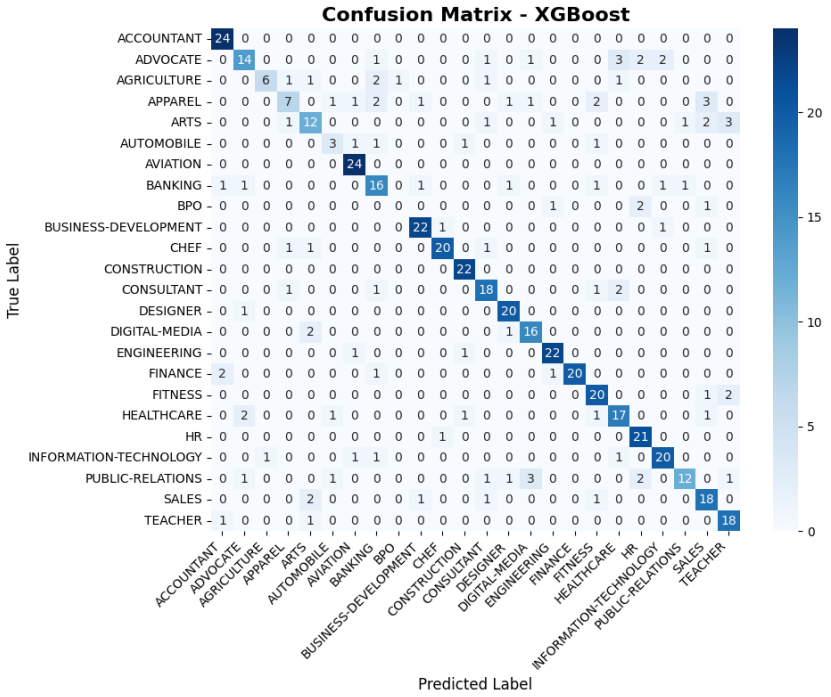


Figure 6. Confusion Matrix for the XGBoost Model. The vertical axis represents the true labels, and the horizontal axis represents the predicted labels. The values on the main diagonal represent the number of correct classifications.

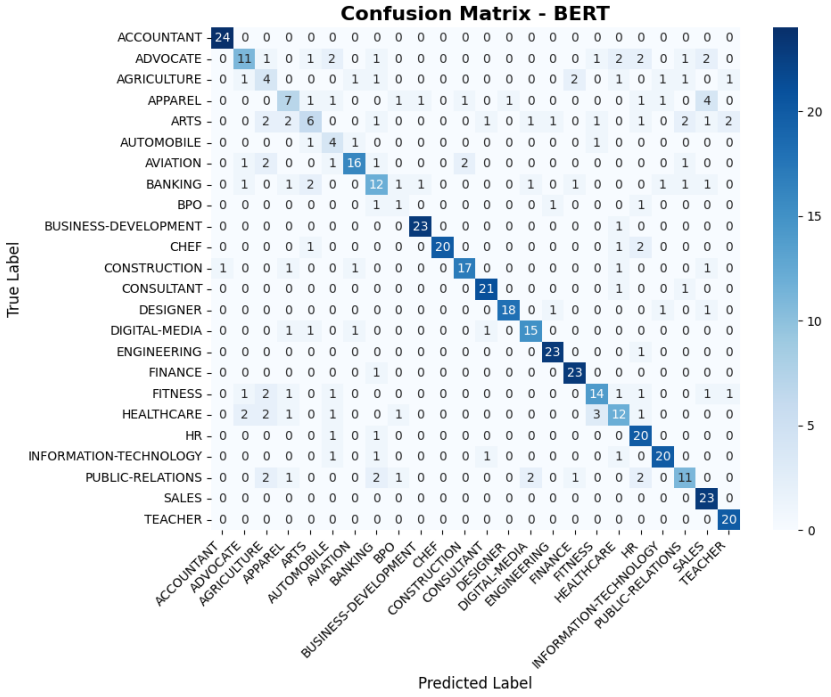


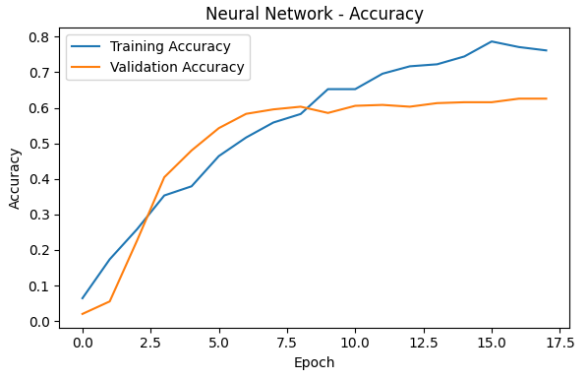
Figure 7. Confusion Matrix for the BERT Model. The main diagonal graph shows the number of correct predictions for each of the 24 job categories

These matrices provide visual evidence of the specific strengths and weaknesses of each approach. The XGBoost confusion matrix (Figure 6) confirms its high performance on well-defined categories, achieving near-perfect accuracy on ACCOUNTANT, AVIATION, and CONSTRUCTION. However, the matrix also reveals logical misclassifications between semantically adjacent categories, such as two FINANCE resumes being incorrectly classified as ACCOUNTANT.

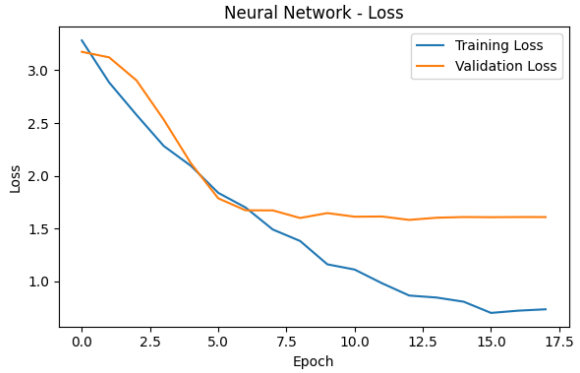
On the other hand, the BERT confusion matrix (Figure 7) highlights its superior contextual understanding. BERT demonstrates better performance on semantically overlapping categories, such as its success in correctly classifying 21 out of 24 CONSULTANT resumes, a significant improvement over XGBoost which only succeeded on 18 samples. This visual analysis reinforces the conclusion that while XGBoost is superior in overall metrics, BERT has an advantage in handling semantic ambiguity between job categories.

3.3 Specific Analysis of Deep Learning Models

The performance of the Neural Network (MLP) model, which achieved a modest F1-score of 0.637, can be better understood by analyzing its training dynamics, as visualised in the learning curve in Figure 8.



(a)



(b)

Figure 8. Learning curves of the Neural Network model over 18 epochs. Plot (a) shows the training and validation accuracy, while plot (b) shows the training and validation loss.

A detailed review of its training process reveals a classic case of overfitting on a sparse dataset. While the training accuracy steadily increased throughout the epochs, the validation accuracy began to plateau around the 60-62% mark. Mechanisms such as ReduceLROnPlateau, which was triggered twice (at epochs 12 and 16) to adjust the learning rate, and EarlyStopping were employed to combat this. The training was ultimately halted at epoch 18, and the model weights from the best-performing epoch (epoch 13) were restored. This behavior underscores the challenge for an MLP architecture, which learns from scratch, to generalize effectively from the limited and high-dimensional TF-IDF features without memorizing the training data.

In stark contrast, the superior performance of BERT (F1-score of 0.728) highlights the significant benefit of transfer learning. Unlike the MLP, which relied on TF-IDF features, the BERT-based model leveraged a pre-trained transformer architecture to generate rich, contextual embeddings for each resume. This process of transfer learning allows the model to apply a deep, pre-existing understanding of language, grammar, and semantic relationships to the classification task. The model is not learning from a blank slate but is instead fine-tuning its vast linguistic knowledge for the specific domain of resume screening. The significant performance gap between the two models therefore illustrates the profound advantage of transfer learning over learning from scratch, especially in NLP tasks with moderately sized, domain-specific datasets where contextual nuance is critical.

3.4 Implications and Further Discussion

The findings from this research present a complex landscape of trade-offs between predictive performance, interpretability, and computational requirements in the context of resume screening. XGBoost is definitively the optimal choice for achieving the highest predictive accuracy (F1-score of 0.779), making it the primary candidate for applications where performance is the sole priority. However, in the Human Resources domain, where decisions can significantly impact individuals' careers, interpretability is not merely an additional feature but a necessity. This need is driven by ethical and legal considerations, such as ensuring a fair recruitment process, avoiding discrimination, and complying with regulations like the GDPR's "right to an explanation."

This presents a practical dilemma. Random Forest, with its strong F1-score (0.711), offers far greater transparency through its ability to extract feature importances. This allows HR practitioners to understand the basis of a model's decision, for instance, which keywords or phrases were most influential in a classification. Organizations must therefore consider whether the approximate 7% F1-score performance gain offered by XGBoost is worth the risk of using a "black-box" model that is difficult to audit. For many companies, particularly those in highly regulated industries, the "good enough" performance of Random Forest, coupled with its procedural transparency, may represent a more responsible and legally defensible choice.

This discussion is further enriched by the performance of BERT, which introduces a third dimension to the trade-off: semantic capability versus computational cost. With an F1-score (0.728) that surpasses Random Forest, BERT proves to be a powerful option when the classification task is highly dependent on understanding contextual nuances that keyword-based models might miss. An example includes distinguishing between roles

with overlapping job descriptions. However, this semantic power comes at a significant computational cost for both training and inference, making it a more suitable choice for large enterprises with mature AI infrastructure.

Ultimately, while model selection is important, this study conclusively demonstrates that the primary limitation lies within the data itself. The universally poor performance on minority classes, such as the near-zero F1-score for BPO, is the starkest evidence of this. This finding also highlights the effectiveness and limitations of the adaptive imbalance handling strategy. Although advanced techniques like SMOTE and SMOTEENN were designated for more severe cases, the computational constraints of sparse data necessitated a fallback to the class weights method. The subsequent failure of models to learn patterns from extremely underrepresented classes indicates that even this robust, adaptive approach has its limits when faced with severe data scarcity. This partial success underscores that these methods are not a panacea and reinforces the argument that the fundamental challenge lies in the data's representativeness. Therefore, future research should prioritize the following areas: exploring advanced data augmentation techniques, such as back-translation or generative models to create quality synthetic samples; investigating few-shot or zero-shot learning approaches; and conducting explicit fairness and bias audits to measure and mitigate algorithmic biases.

4. Conclusions

This study conducted a systematic comparative evaluation of seven machine learning models on an imbalanced resume dataset, finding that XGBoost (F1-score of 0.779) delivered the highest predictive performance. The study's main contribution is its analysis of the multi-dimensional trade-off facing HR practitioners: the choice between maximum performance (XGBoost), high interpretability for transparency and audits (Random Forest, F1-score of 0.711), or deep semantic understanding at a higher computational cost (BERT, F1-score of 0.728).

The findings also starkly highlight data scarcity as the most significant barrier, as performance on minority classes remained poor despite the application of an adaptive imbalance handling strategy. This limitation, along with the study's moderate dataset size, directs future research to shift from model comparisons to more data-centric approaches. Promising directions include exploring generative data augmentation techniques, investigating few-shot learning models, and integrating fairness and bias audit frameworks to ensure that automated recruitment systems are developed to be not only accurate but also equitable and responsible.

Acknowledgements

The authors declare no conflict of interest. This research was funded by Lembaga Penelitian dan Pengabdian Masyarakat Universitas Airlangga trough "International Research Collaboration TOP #300 (grant number 339/UN3.15/PT/2023), and author gratefully acknowledge Febdian Rusydi (Physics, Universitas Airlangga, Indonesia) for his constructive feedback during manuscript preparation.

References

[1] A. K. Sinha, Md. Amir Khusru Akhtar, and A. Kumar, "Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review," 2021, pp. 207–214. doi: 10.1007/978-981-33-4859-2_21.

[2] Dr. Sandeep Tayal, T. Sharma, S. Singhal, and A. Thakur, "Resume Screening using Machine Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 602–606, Apr. 2024, doi: 10.32628/CSEIT2410275.

[3] C. Qin *et al.*, "A Comprehensive Survey of Artificial Intelligence Techniques for Talent Analytics," *Proceedings of the IEEE*, vol. 113, no. 2, pp. 125–171, Feb. 2025, doi: 10.1109/JPROC.2025.3572744.

[4] A. S. Imran, H. Hodnefjeld, Z. Kastrati, N. Fatima, S. M. Daudpota, and M. A. Wani, "Classifying European Court of Human Rights Cases Using Transformer-Based Techniques," *IEEE Access*, vol. 11, pp. 55664–55676, 2023, doi: 10.1109/ACCESS.2023.3279034.

[5] M. Maree and W. Shehada, "Optimizing Curriculum Vitae Concordance: A Comparative Examination of Classical Machine Learning Algorithms and Large Language Model Architectures," *AI*, vol. 5, no. 3, pp. 1377–1390, Aug. 2024, doi: 10.3390/ai5030066.

[6] R. Sharma, A. Rajput, and V. K. Srivastav, "Job Description and Resume Matching System Using NLP," in *2025 International Conference on Next Generation Information System Engineering (NGISE)*, IEEE, Mar. 2025, pp. 1–5. doi: 10.1109/NGISE64126.2025.11085197.

[7] M. Liu, C. Guo, and L. Xu, "An interpretable automated feature engineering framework for improving logistic regression," *Appl Soft Comput*, vol. 153, p. 111269, Mar. 2024, doi: 10.1016/j.asoc.2024.111269.

[8] P. Skondras, P. Zervas, and G. Tzimas, "Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification," *Future Internet*, vol. 15, no. 11, p. 363, Nov. 2023, doi: 10.3390/fi15110363.

[9] A. Heakl, Y. Mohamed, N. Mohamed, A. Elsharkawy, and A. Zaky, "ResumeAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models," *Procedia Computer Science 244*, pp. 158–165, Jul. 2024., https://doi.org/10.48550/arXiv.2406.18125

[10] P. Yang and J. Li, "Construction and application of intelligent human resource management system based on machine learning algorithm," *Journal of Computational Methods in Sciences and Engineering*, vol. 25, no. 4, pp. 3684–3696, Jul. 2025, doi: 10.1177/14727978251318814.

[11] M. Haddouchi and A. Berrado, "Forest-ORE: Mining an optimal rule ensemble to interpret random forest models," *Eng Appl Artif Intell*, vol. 143, p. 109997, Mar. 2025, doi: 10.1016/j.engappai.2024.109997.

[12] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: a comparative study of machine learning models," *J Reliab Intell Environ*, vol. 11, no. 1, p. 1, Mar. 2025, doi: 10.1007/s40860-024-00240-0.

[13] B. V. N. Reddy and S. Parthiban, "Recognizing the required and skill-based resume to the companies with BERT compared with random forest," 2025, p. 020155. doi: 10.1063/5.0262799.

[14] A. Deshmukh and A. Raut, "Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking," *Annals of Data Science*, vol. 12, no. 2, pp. 591–603, Apr. 2025, doi: 10.1007/s40745-024-00524-5.

[15] S. Regilan, P. Gajalakshmi, D. Weslin, J. Vijay, D. Kadhiravan, and J. Jenitha, "Benchmarking AI-Driven Resume Screening: an Evaluation of Precision and Efficiency," in *2025 11th International Conference on Communication and Signal Processing (ICCSP)*, IEEE, Jun. 2025, pp. 783–788. doi: 10.1109/ICCSP64183.2025.11089249.

[16] L. Roumbanis, "On the present-future impact of AI technologies on personnel selection and the exponential increase in meta-algorithmic judgments," *Futures*, vol. 166, p. 103538, Feb. 2025, doi: 10.1016/j.futures.2025.103538.

[17] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *Calif Manage Rev*, vol. 61, no. 4, pp. 15–42, Aug. 2019, doi: 10.1177/0008125619867910.

[18] I. Puspitasari, "The impacts of consumer's health topic familiarity in seeking health information online," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, IEEE, Jun. 2017, pp. 104–109. doi: 10.1109/SERA.2017.7965714.

[19] I. Puspitasari, A. N. R. Wiambodo, and P. Soeparman, "The impact of expectation confirmation, technology compatibility, and customer's acceptance on e-wallet continuance intention," 2021, p. 050012. doi: 10.1063/5.0042269.

[20] D. Vrontis, M. Christofi, V. Pereira, S. Tarba, A. Makrides, and E. Trichina, "Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review," *The International Journal of Human Resource Management*, vol. 33, no. 6, pp. 1237–1266, Mar. 2022, doi: 10.1080/09585192.2020.1871398.

[21] R. Govindarajan, N. K. Kumar, S. R. P, S. P. E, D. B, and P. K. G, "Predicting Employee Attrition: A Comparative Analysis of Machine Learning Models Using the IBM Human Resource Analytics Dataset," *Procedia Comput Sci*, vol. 258, pp. 4084–4093, 2025, doi: 10.1016/j.procs.2025.04.659.

[22] I. Ajunwa, "The 'black box' at work," *Big Data Soc*, vol. 7, no. 2, Jul. 2020, doi: 10.1177/2053951720938093.

[23] Y. Zhuang, "An Improved TFIDF Algorithm in Electronic Information Feature Extraction Based on Document Position," 2012, pp. 449–454. doi: 10.1007/978-3-642-31516-9_72.

[24] Y. Zhang, L. Gong, and Y. Wang, "An improved TF-IDF approach for text classification," *J Zhejiang Univ Sci*, vol. 6, no. 1, pp. 49–55, Jan. 2005, doi: 10.1631/jzus.2005.A0049.

[25] Y. A. Alhaj, M. A. A. Al-qaness, A. Dahou, M. Abd Elaziz, D. Zhao, and J. Xiang, "Effects of Light Stemming on Feature Extraction and Selection for Arabic Documents Classification," 2020, pp. 59–79. doi: 10.1007/978-3-030-34614-0_4.

[26] P. Sheridan and M. Onsjö, "The hypergeometric test performs comparably to TF-IDF on standard text analysis tasks," *Multimed Tools Appl*, vol. 83, no. 10, pp. 28875–28890, Sep. 2023, doi: 10.1007/s11042-023-16615-z.

[27] S. Guo, "Application and Impact of Position Embedding in Natural Language Processing," in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, IEEE, Apr. 2024, pp. 246–250. doi: 10.1109/ICEIB61477.2024.10602658.

[28] E. Qais and V. MN, "Transforming Qualitative Business Decisions through Text Analytics: An NLP Approach," *Int J Res Appl Sci Eng Technol*, vol. 13, no. 1, pp. 910–918, Jan. 2025, doi: 10.22214/ijraset.2025.66504.

[29] Ncr and J. Clinton, "CRISP-DM 1.0 Step-by-step data mining guide," DaimlerChrysler, 1999, https://api.semanticscholar.org/CorpusID:59777418

[30] S. Colin, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing*, vol 5, pp. 13-22, 2000.

[31] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[32] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf Process Manag*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[34] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[36] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[37] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[38] N. Ravikumar, A. Zakeri, Y. Xia, and A. F. Frangi, "Deep learning fundamentals," in *Medical Image Analysis*, Elsevier, 2024, pp. 415–450. doi: 10.1016/B978-0-12-813657-7.00041-8.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186. Association for Computational Linguistics, May 2019, https://doi.org/10.18653/v1/N19-1423

[40] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352–356, Apr. 2023, doi: 10.47852/bonviewJCCE3202838.

[41] M. A. Amirabadi, M. H. Kahaei, and S. A. Nezamalhosseini, "Novel suboptimal approaches for hyperparameter tuning of deep neural network [under the shelf of optical communication]," *Physical Communication*, vol. 41, p. 101057, Aug. 2020, doi: 10.1016/j.phycom.2020.101057.

[42] A. M. Shetty, M. F. Aljunid, D. H. Manjaiah, and A. M. S. Shaik Afzal, "Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis," 2024, pp. 451–474. doi: 10.1007/978-981-99-7814-4_36.

[43] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Comput Stat*, vol. 36, no. 3, pp. 2009–2031, Sep. 2021, doi: 10.1007/s00180-020-00999-9.

[44] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 2229-398, Oct. 2020, https://doi.org/10.48550/arXiv.2010.16061

[45] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.01406116.

[46] E. Mortaz, "Imbalance accuracy metric for model selection in multi-class imbalance classification problems," *Knowl Based Syst*, vol. 210, p. 106490, Dec. 2020, doi: 10.1016/j.knosys.2020.106490.

[47] A. Hussain and R. Rizwan, "Strategic AI adoption in SMEs: A Prescriptive Framework," Aug. 2024, https://doi.org/10.48550/arXiv.2408.11825

[48] F. A. Najafabadi, J. Bogner, I. Gerostathopoulos, and P. Lago, "An Analysis of MLOps Architectures: A Systematic

Mapping Study," Jun. 2024, doi: 10.1007/978-3-031-70797-1_5.

[49] A. Fabris *et al.*, "Fairness and Bias in Algorithmic Hiring: a Multidisciplinary Survey," *ACM Transactions on Intelligent Systems and Technology*, vol 16, no. 1, pp.1–54, Jun. 2025, doi: 10.1145/3696457.

[50] M. Chaudhary, L. Gaur, A. Chakrabarti, G. Singh, P. Jones, and S. Kraus, "An integrated model to evaluate the transparency in predicting employee churn using explainable artificial intelligence," *Journal of Innovation & Knowledge*, vol. 10, no. 3, p. 100700, May 2025, doi: 10.1016/j.jik.2025.100700.

[51] Y. Rong *et al.*, "Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 4, pp. 2104–2122, Apr. 2024, doi: 10.1109/TPAMI.2023.3331846.

[52] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies (Basel)*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.

[53] A. Habous and E. H. Nfaoui, "Combining Word Embeddings and Deep Neural Networks for Job Offers and Resumes Classification in IT Recruitment Domain," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021, doi: 10.14569/IJACSA.2021.0120774.