

A Machine Learning Model for Bulk Resume Classification and Recommendation for Human Resource Managers

Kennedy Ngetich
Department of Computer
Science and Information
Technology
Kabarak University
Nakuru, Kenya

Ruth Oginga
Department of Computer
Science and Information
Technology Kabarak
University
Nakuru, Kenya

Nelson Masese
Department of Computer Science
and Information Technology
Kabarak University
Nakuru, Kenya

Abstract: Manual resume screening does not scale to the volume of applications a single vacancy now attracts, while conventional Applicant Tracking Systems rely on rigid keyword matching, evaluate one job domain at a time and offer no explanation for their decisions. This study designed, implemented and evaluated a machine learning model for bulk resume classification and recommendation, integrating a fine-tuned BERT classifier, LLaMA2- and Flan-T5-based recommendation generation and a LIME explainability layer within an ethical governance framework. A Design Science Research methodology, combined with Design Thinking and an Agile life cycle, was executed across six two-week sprints. The dataset comprised 14,450 labelled resumes spanning five job domains, split 70/15/15. The fine-tuned classifier achieved 91.3% accuracy and a macro-averaged F1-score of 0.893 on 2,168 held-out resumes, outperforming Support Vector Machine, Random Forest and Naïve Bayes baselines by 8.8 to 16.5 percentage points. A Demographic Parity Ratio audit across gender returned 0.96–1.04 in every domain. User Acceptance Testing with 30 Human Resource professionals in Nakuru County, Kenya returned Technology Acceptance Model scores of 4.30, 4.17 and 4.23 for perceived usefulness, ease of use and fairness, each above the 4.0 adoption-intent threshold.

Keywords: resume classification; BERT; explainable artificial intelligence; algorithmic fairness; human resource management; large language models

1. INTRODUCTION

The Society for Human Resource Management estimated that a single corporate job posting attracted an average of 250 applications by 2023 [1], and Applicant Tracking Systems (ATS) had, by the early 2020s, been adopted by an estimated 98% of Fortune 500 companies and a growing share of small and medium enterprises [2]. Manually reviewing 250 applications at approximately five minutes per resume would demand over twenty hours of dedicated screening time, a figure broadly consistent with the industry time-burden literature rather than drawn from a single verified study. Under this volume pressure recruiters frequently default to unconscious pattern-matching that favours candidates whose names, educational institutions or personal characteristics match implicit stereotypes, even where explicit anti-discrimination policy exists [3], [4]. In Kenya specifically, bias in public-sector hiring has been linked to patterns of nepotism and ethnic favouritism that undermine organisational performance and public trust [5].

This paper reports the design, implementation and evaluation of an open-source, multi-domain resume classification and recommendation model built for a resource-constrained African recruitment context. The remainder of the paper surveys the relevant literature, defines the problem and objectives addressed, sets out the methodology, reports and discusses the results, and identifies the scope for further work.

2. LITERATURE SURVEY

2.1 Resume Classification Approaches

Prior resume-classification research has evolved from classical machine learning toward transformer-based architectures. Pal et al. [6] and Ali et al. [7] evaluated Naïve Bayes, Support Vector Machines, Random Forest and K-Nearest Neighbours for resume classification in controlled experimental settings, reporting moderate accuracy but limited capacity to capture contextual dependency between words. Kaur and Gupta [8] applied fine-tuned BERT to resume classification and reported F1-scores between 0.87 and 0.91, a benchmark within which this study's own result ($F1 = 0.893$) sits. Saatçı et al. [9] demonstrated that Natural Language Processing-based competency extraction reduces both screening time and bias relative to keyword-based filtering, while Gan et al. [10] showed that Large Language Model (LLM)-powered screening agents can operate up to eleven times faster than human reviewers while maintaining approximately 88% F1-score on competency matching.

2.2 Transformer Architectures and Variants

Since Devlin et al.'s original release of BERT [11], several variants have emerged. RoBERTa [12] removes BERT's Next Sentence Prediction objective and trains on more data with dynamic masking; DistilBERT [13] compresses BERT via knowledge distillation to roughly 60% of its parameter count while retaining approximately 97% of its language understanding; and ALBERT [14] reduces memory

consumption through cross-layer parameter sharing. Domain-adapted variants such as BioBERT [15] and SciBERT [16] demonstrate that BERT's architecture can be specialised to a target domain's vocabulary, a strategy conceptually related to the vocabulary augmentation for Kenyan-context resume text adopted here. More recent architectures, including DeBERTa [17] and ELECTRA [18], have outperformed the original BERT on several benchmark tasks; bert-base-uncased was nonetheless retained as this study's classification backbone on account of its extensively validated precedent in resume classification specifically [8], its first-class support within the Hugging Face Transformers ecosystem, and its favourable balance of reliability and computational cost. Sentence-BERT [19], which produces sentence-level embeddings suited to semantic similarity and retrieval tasks, is identified as a technically justified candidate for a related but distinct downstream use – matching a resume against a job description at scale – rather than the five-class domain classification task addressed here.

2.3 Algorithmic Fairness

Barocas et al. [20] established that several statistical definitions of algorithmic fairness are mutually incompatible, so that no single metric can simultaneously satisfy every intuitive notion of fair treatment; this study accordingly adopted Demographic Parity as its explicit, stated fairness criterion. Wilson and Caliskan [4] extended this concern to intersectional bias in LLM-driven hiring specifically, a consideration that motivated the explicit scoping decision taken here to evaluate gender and educational-background parity without extending to a full intersectional analysis.

2.4 Theoretical Framework

Three theories jointly framed this study. Information Processing Theory [21] models cognition as encoding, storage and retrieval, mapped here onto the classification pipeline's read–learn–decide structure; a recognised critique of the theory, that it treats cognition as a linear process without accounting for affective dimensions such as recruiter fatigue, is precisely what motivates positioning the model as an augmentation of human judgement rather than a replacement for it. The Technology Acceptance Model (TAM) [22] holds that perceived usefulness and perceived ease of use jointly determine technology adoption; this study extended the model with a third construct, perceived fairness, following the argument that fairness is an increasingly critical determinant of AI adoption in HR contexts specifically [23]. Algorithmic Fairness Theory [20] supplied the study's operational definition of fairness and the Demographic Parity Ratio (DPR) metric used in its bias audit. Together these three theories address a fragmentation the literature review found common in prior resume-classification research, where studies typically reported strong performance metrics without a corresponding fairness or adoption analysis, or vice versa.

3. PROBLEM DEFINITION

Applicant Tracking Systems, designed to automate initial screening, are themselves constrained. Their keyword dependency systematically excludes candidates who describe equivalent competencies using synonymous but non-matching terminology [24], and commercial platforms such as Workday and iCIMS carry per-seat subscription costs that place them beyond the reach of most Kenyan small and medium enterprises [25]. They are further limited by single-domain configuration, which prevents simultaneous evaluation across multiple job fields, and by an absence of explainability, which

prevents HR managers from identifying and correcting biased filtering decisions [26].

No documented open-source, multi-domain, LLM-powered resume classification system integrating bias mitigation, explainability and contextual language understanding was identified in the reviewed literature for deployment in resource-constrained African recruitment environments. This gap in both knowledge and practice – specifically, the absence of such a system designed and evaluated within the Kenyan recruitment context – is the central problem this study addressed.

The general objective was to design a machine learning model for bulk resume classification and recommendation for human resource managers, comprising two sequential functions: classification of each resume into a job domain by a fine-tuned BERT model, and recommendation of shortlisted candidates through LLaMA2- and Flan-T5-generated, HR-readable rationales. Four specific objectives guided the study:

- i. To identify the challenges facing existing resume screening approaches that inhibit effective bulk candidate evaluation for human resource managers;
- ii. To design a machine learning model for bulk resume classification and recommendation that addresses the identified challenges and requirements;
- iii. To implement the designed model as a functional prototype; and
- iv. To evaluate the model's classification performance, fairness and user acceptance among HR professionals.

4. METHODOLOGY

4.1 Research Design

The study followed a Design Science Research (DSR) methodology [27], combined with a Design Thinking process (Empathise, Define, Ideate, Prototype, Test) for requirements elicitation and prototype refinement, and an Agile Software Development Life Cycle executed across six two-week sprints for implementation. This combination allowed qualitative requirements gathering to inform iterative technical build-out within an overarching artefact-evaluation logic.

4.2 Study Area

The study was conducted within Nakuru County, Kenya, targeting recruitment firms and HR departments across manufacturing, public service, agriculture, retail and professional services. Nakuru County was selected as a regionally under-served but economically active recruitment environment, mirroring the study's broader rationale of targeting an under-served market rather than replicating research already conducted in data-rich Western contexts.

4.3 Samples and Data Collection

Three distinct, non-overlapping human participant samples were used. An Interview Sample ($n = 17$), drawn during the Empathise phase, comprised semi-structured interviews with HR managers and recruitment officers conducted to elicit functional requirements. A Pilot Sample ($n = 5$), drawn from an organisation deliberately excluded from the main evaluation sample to prevent double participation, validated the prototype and research instrument, returning a Cronbach's alpha of 0.87. A User Acceptance Testing (UAT) Sample ($n = 30$), composed of 12 HR Managers, 10 Recruitment Officers and 8 Hiring Personnel, evaluated the completed prototype's usability, performance and fairness. The UAT sample size

was justified not by a probability-sampling formula but by the recommended minimum of 10–20 domain experts for Design Science Research artefact evaluation [28].

The classification dataset comprised 12,450 resumes scraped across five sources – Kaggle Resume Datasets, GitHub open resume repositories, Indeed public job listings, LinkedIn public profiles and the LiveCareer resume examples portal – augmented with 2,000 synthetically generated, Kenyan-context resumes to correct underrepresentation of local resume formatting and terminology. This produced 14,450 total labelled records across five job domains (Software Engineering, Data Science, Human Resources, Sales and Marketing, and Finance), divided using a stratified 70/15/15 train/validation/test split (10,115 / 2,168 / 2,168 records respectively).

4.4 Model Architecture and Preprocessing

Resume text was cleaned, imputed for missing structured fields and processed through spaCy's `en_core_web_sm` Named Entity Recognition model, fine-tuned on annotated resumes, before WordPiece tokenisation for BERT. Kenyan resumes frequently exhibit English–Swahili code-switching within role titles (for example, “Mkurugenzi wa Masoko / Marketing Director”); this was addressed through a targeted vocabulary augmentation step adding the 500 most frequent Swahili terms observed in the training corpus to the tokenizer's vocabulary, acknowledged as a partial rather than complete mitigation given that a full bilingual pre-training run was outside the study's compute budget.

`bert-base-uncased` was fine-tuned with a softmax classification head over the five job-domain categories, using the AdamW optimizer with linear learning-rate warmup, a maximum sequence length of 512 tokens, batch size 16, learning rate 2×10^{-5} and weight decay 0.01, on a GPU-accelerated Google Colab Pro+ environment. Training was scheduled for 8 epochs but converged at epoch 6 (validation accuracy 90.8%) under early stopping (patience = 3) once validation loss began to plateau while training loss continued to fall, an overfitting safeguard rather than a resource-saving shortcut. Class imbalance across the five domains and across gender within domain was addressed using the Synthetic Minority Oversampling Technique (SMOTE, $k = 5$), applied in preference to random undersampling (which would have discarded valid Finance and Software Engineering examples) or naive oversampling (which risks overfitting).

Following classification, resumes scoring above a defined confidence threshold were passed to a recommendation layer combining a 7-billion-parameter LLaMA2 model [29], which extracts structured key points from the resume text, and Flan-T5 [30], which composes the plain-language recommendation narrative presented to the HR user; both operate as a complementary layer to, rather than a replacement for, the BERT classification decision. Explainability was provided through LIME [31], configured to surface up to eight word-level features per classification, presented to the HR user as matched and missing keywords relative to the target domain.

4.5 Evaluation Approach

Classification performance was benchmarked against Support Vector Machine (RBF kernel), Random Forest (200 trees) and Naïve Bayes baselines using accuracy and macro-averaged F1-score against pre-defined thresholds of 85% and 0.80 respectively, with statistical significance assessed via McNemar's test. Fairness was audited using the Demographic

Parity Ratio across gender, with an acceptable range conventionally set at 0.80–1.25. User acceptance was measured via a 13-item, 5-point Likert-scale instrument (Cronbach's $\alpha = 0.87$ overall) operationalising three TAM constructs – Perceived Usefulness (5 items), Perceived Ease of Use (3 items) and Perceived Fairness (2 items) – with a further 3-item scale evaluating explanation quality (Clarity, Relevance, Actionability), disaggregated by respondent role. Ethical governance measures included full anonymisation of resume data prior to processing, encrypted storage, compliance with Kenya's Data Protection Act (2019), and an explicit human-override safeguard ensuring no participating organisation's recruitment outcome during the study period was determined by the prototype's output alone.

5. RESULTS AND DISCUSSION

5.1 Functional Requirements (Objective 1)

Analysis of the 17 Empathise-phase interviews identified ten functional requirements for an effective bulk resume screening tool, tiered by the proportion of respondents who volunteered each requirement unprompted. The four Critical/High-priority requirements were multi-domain simultaneous classification (100%), plain-language explanation of classification decisions (88%), demographic bias and fairness reporting (82%), and automated candidate ranking within domain (76%). These four requirements directly informed the classification, explainability and ranking architecture of the resulting prototype; the remaining Medium/Low-priority requirements – skill-gap highlighting, sub-5-second processing, non-technical dashboard accessibility and a model-retraining feedback loop – received correspondingly lighter validation attention during implementation.

5.2 Model Performance (Objective 2)

On the held-out test set of 2,168 resumes, the fine-tuned BERT classifier achieved 91.3% accuracy and a macro-averaged F1-score of 0.893, exceeding both pre-defined thresholds (85% accuracy; 0.80 macro-F1). As shown in Figure 1, this exceeded the strongest classical baseline, Support Vector Machine (82.1% accuracy, $F1 = 0.805$), by 8.8 percentage points in F1, and exceeded Random Forest (80.4%, $F1 = 0.784$) and Naïve Bayes (74.8%, $F1 = 0.728$) by wider margins, with McNemar's test confirming the BERT-versus-SVM gap as statistically significant ($p < .001$).

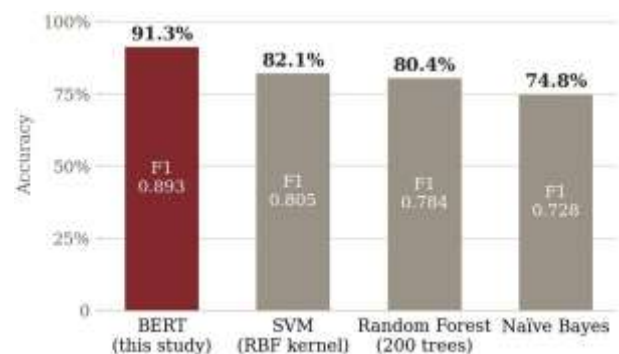


Figure 1. Classification accuracy and macro-averaged F1-score of the fine-tuned BERT classifier against classical machine learning baselines

Per-domain performance ranged from 89.8% accuracy (Sales and Marketing) to 93.1% (Software Engineering), as shown in

Figure 2. The spread is attributed to Software Engineering and Data Science resumes using a comparatively narrow, well-documented technical vocabulary, against Sales and Marketing's far wider range of role titles and organisation-specific jargon.



Figure 2. Classification accuracy by job domain on the held-out test set (n = 2,168)

5.3 Prototype Implementation (Objective 3)

The completed prototype, named Talensis, comprises a resume-upload and preprocessing pipeline feeding a fine-tuned BERT classifier, whose output is passed to a parallel downstream layer combining LLaMA2-based key-point extraction, Flan-T5-based recommendation generation and LIME-based explainability, before rendering ranked, explained and fairness-audited results on an HR-facing dashboard. Figure 3 illustrates this end-to-end architecture, including the periodic-retraining feedback loop connecting dashboard output back into the classification stage.

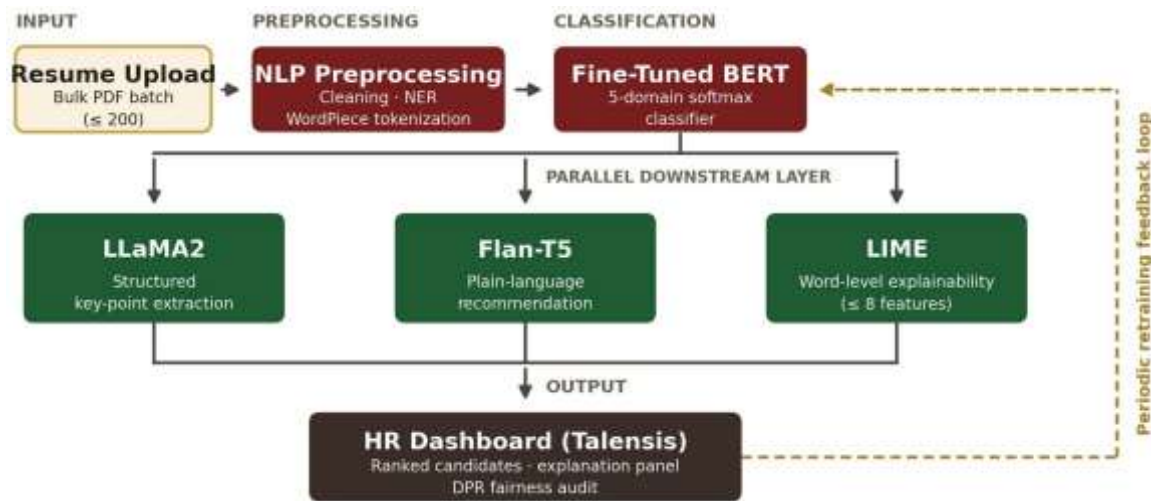


Figure 3. End-to-end system architecture of the Talensis prototype

From the HR user's perspective this architecture is presented as a straightforward five-stage workflow – Upload, Select, Classify, Review and Audit – shown in Figure 4, directly

operationalising the top functional requirements identified under Objective 1.



Figure 4. The five-stage Talensis user workflow, from bulk upload to fairness-audited output

The dashboard supported batches of up to 200 resumes with a median per-resume response time of 3.8 seconds, achieved through batch-processed BERT inference (improved from an initial pilot figure of 6.2 seconds). During UAT, 90% of participants completed assigned tasks without facilitator assistance; the remaining 10% requiring assistance were, without exception, participants reporting no prior experience with any Applicant Tracking System, suggesting the assistance rate tracked prior digital-tool familiarity rather than a prototype usability defect.

5.4 Fairness and User Acceptance (Objective 4)

The Demographic Parity Ratio audit across gender returned values of 0.96–1.04 across all five job domains, within the conventionally accepted 0.80–1.25 range; a comparable education-level (diploma versus degree) audit likewise fell within range. The TAM-based UAT results returned mean scores of 4.30 (SD = 0.82) for Perceived Usefulness, 4.17 (SD = 0.79) for Perceived Ease of Use and 4.23 (SD = 0.83) for Perceived Fairness, each exceeding the study's 4.0 adoption-intent threshold on a 5-point scale. Explanation-quality ratings, disaggregated by role, showed a consistent gradient: HR Managers rated explanation usefulness highest, followed

by Recruitment Officers, then Hiring Personnel – a pattern plausibly linked to HR Managers' greater accountability for final shortlisting decisions, though confirming this would require a dedicated qualitative follow-up study.

5.5 Discussion

The classification results are consistent with, and at the upper end of, the 0.87–0.91 F1-score range reported for fine-tuned BERT resume classification [8], and the magnitude of BERT's advantage over classical baselines tracks closely with each algorithm's theoretical capacity to capture contextual dependency: none in Naïve Bayes's independence assumption, partial in tree- and margin-based classifiers operating on static term-frequency features, and full bidirectional context in BERT's transformer encoder. This pattern supports the broader literature's shift away from classical machine learning toward transformer architectures for text classification tasks of this kind.

The fairness result (DPR 0.96–1.04) indicates that the combination of SMOTE-based class and demographic rebalancing during training and explicit DPR auditing at deployment is a workable, if not sufficient, response to the fairness-incompatibility concern raised by Barocas et al. [20]: this study makes no claim to have satisfied every statistical fairness definition, only the specific Demographic Parity criterion it explicitly adopted. The user acceptance findings extend the argument for perceived fairness as a third TAM construct in AI-in-HR contexts [23], and the observed role-based gradient in explanation-quality ratings offers a plausible, though not yet confirmed, refinement: that perceived explanation quality may itself be moderated by a respondent's proximity to final hiring accountability, a novel finding relative to the reviewed TAM literature, which treated perceived fairness as a single undifferentiated construct.

6. CONCLUSION

This study designed, implemented and evaluated a machine learning model for bulk, multi-domain resume classification and recommendation, integrating a fine-tuned BERT classifier, LLM-based recommendation generation, LIME-based explainability and an explicit fairness audit within an ethical governance framework. Across all four specific objectives the model met or exceeded its pre-defined thresholds: classification accuracy and F1-score exceeded both the target thresholds and the classical baselines by a substantial and statistically significant margin; the resulting prototype was usable by HR professionals with minimal assistance; and the fairness audit and user acceptance scores each cleared their respective thresholds. HR practitioners in the User Acceptance Testing sample did not merely tolerate the resulting system but actively endorsed both its usefulness and its fairness safeguards – evidence that transformer-based classification, when paired with explainability and fairness auditing rather than deployed as an opaque filter, can gain practitioner trust in a resource-constrained African recruitment context.

The study contributes theoretically through the integration of Information Processing Theory, the Technology Acceptance Model extended with a Perceived Fairness construct, and Algorithmic Fairness Theory into a single evaluation framework, addressing a fragmentation in prior resume-classification literature between performance-focused and fairness- or adoption-focused studies. Methodologically, it contributes a replicable three-sample evaluation design separating qualitative requirements elicitation from quantitative acceptance and fairness testing. Empirically, it

contributes a validated benchmark for transformer-based resume classification, fairness auditing and practitioner acceptance within an East African recruitment context, together with a documented, locally grounded mitigation for English–Swahili code-switching in resume text. Practically, it contributes an open-source, low-infrastructure deployment profile directly addressing the affordability barrier that limits commercial ATS adoption among Kenyan small and medium enterprises.

For practice, organisations managing application volumes exceeding 100 resumes per vacancy are best positioned to benefit from a model of this kind. HR departments should retain the LIME explanation-panel output – matched keywords and skill-gap rationale – as part of the recruitment audit trail, providing a defensible justification for shortlisting decisions should they later be questioned by an unsuccessful applicant or a regulatory body, and, given the observed role-based gradient, should differentiate training on interpreting model explanations by role. For policy, the findings motivate, without their own establishing, a regulatory framework for algorithmic hiring tools developed jointly by the National Commission for Science, Technology and Innovation and the Kenya National Human Rights Commission, a public-sector pilot by the Ministry of Public Service and Human Capital Development, and the incorporation of AI-screening literacy, including interpretation of fairness audit reports, into Institute of Human Resource Management continuing professional development curricula.

7. FUTURE SCOPE

Several limitations qualify these findings and define the scope for further work. The 2,000 synthetic resumes were present across all three data splits rather than the training split alone, meaning the reported 91.3% accuracy may somewhat overstate real-world generalisation performance; a synthetic-free supplementary test-set evaluation is recommended as a priority refinement before this figure is relied upon as a real-world estimate, and no formal data-leakage check was performed between splits. The mechanism by which each resume's gender attribute was determined for the DPR audit is not detailed in the underlying study documentation, a gap acknowledged rather than a settled methodological detail. The fixed 512-token input limit of bert-base-uncased meant that unusually long, multi-page resumes – disproportionately common among senior academic and research candidates – were truncated, a genuine information loss not separately quantified here. Finally, the UAT panel, though appropriately sized for Design Science Research evaluation, was drawn exclusively from Nakuru County, limiting the generalisability of the acceptance findings to other Kenyan regions or East African labour markets without further replication.

Two further directions are identified. First, a hybrid modelling approach – combining the fine-tuned BERT classifier with rule-based post-processing, ensemble voting across transformer variants, or Sentence-BERT-based retrieval-augmented matching [19] – could plausibly improve both classification robustness and recommendation quality beyond a single fine-tuned model. Second, future work should explicitly investigate the model's behaviour on AI-generated or AI-assisted resumes, including whether such resumes can be reliably detected prior to classification and whether their presence in an applicant pool distorts classification confidence, match scoring or fairness outcomes relative to human-authored resumes – an increasingly pressing extension given the growing availability of AI writing tools to job applicants. Additional priorities include adversarial-

robustness testing against deliberately keyword-stuffed resumes, an explicit intersectional fairness analysis combining gender with age or ethnicity, and replication of this study's evaluation design in other Kenyan counties and East African labour markets.

8. REFERENCES

- [1] CVviz. 2023. The state of recruitment: application volume trends. CVviz Industry Report.
- [2] Coon, A. 2023. ATS adoption trends among Fortune 500 companies. Industry Report.
- [3] Li, L., Li, X., and Lu, Y. 2023. Implicit bias in recruiter decision-making: a review. *J. Organ. Behav.*
- [4] Wilson, C. and Caliskan, A. 2025. Gender, race, and intersectional bias in resume screening via large language models. *arXiv preprint*.
- [5] Mochoge, B., et al. 2015. Nepotism, ethnic favouritism and public sector performance in Kenya. *Afr. J. Public Adm.*
- [6] Pal, R., Shaikh, S., Satpute, S., and Bhagwat, S. 2022. Resume classification using various machine learning algorithms. *ITM Web Conf.* 44, 03011. doi:10.1051/itmconf/20224403011
- [7] Ali, I., Mughal, N., Khan, Z. H., Ahmed, J., and Mujtaba, G. 2024. Resume classification system using natural language processing and machine learning techniques. *Mehran Univ. Res. J. Eng. Technol.* 41, 1, 1-14.
- [8] Kaur, R. and Gupta, S. 2022. Resume screening using BERT: a transfer learning approach. *Int. J. Comput. Appl.*
- [9] Saatçi, B., Kaya, T., and Ünlü, H. 2024. NLP-driven competency extraction for bias-reduced resume screening. *Comput. Human Behav.* 152, 108082. doi:10.1016/j.chb.2024.108082
- [10] Gan, C., Zhang, Q., and Mori, T. 2024. Application of LLM agents in recruitment: a novel framework for resume screening. *arXiv preprint*.
- [11] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, Minneapolis, 4171-4186.
- [12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. 2019. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv:1907.11692*.
- [13] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*.
- [14] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. 2019. ALBERT: a lite BERT for self-supervised learning of language representations. *arXiv:1909.11942*.
- [15] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4, 1234-1240. doi:10.1093/bioinformatics/btz682
- [16] Beltagy, I., Lo, K., and Cohan, A. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP 2019*. Association for Computational Linguistics, Hong Kong, 3615-3620. doi:10.18653/v1/D19-1371
- [17] He, P., Liu, X., Gao, J., and Chen, W. 2021. DeBERTa: decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR 2021)*.
- [18] Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR 2020)*.
- [19] Reimers, N. and Gurevych, I. 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*. Association for Computational Linguistics, Hong Kong, 3982-3992. doi:10.18653/v1/D19-1410
- [20] Barocas, S., Hardt, M., and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. Available from: <http://www.fairmlbook.org>
- [21] Newell, A. and Simon, H. A. 1972. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.
- [22] Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 3, 319-340.
- [23] van Esch, P., Black, J. S., and Ferolie, J. 2019. Marketing AI recruitment: the next phase in job application and selection. *Comput. Human Behav.* 90, 215-222.
- [24] Porebski, P. 2023. The benefits and disadvantages of applicant tracking systems. 4 Corner Resources.
- [25] Mureithi, S. and Kamau, J. 2022. Cost barriers to HR technology adoption among Kenyan SMEs. *East Afr. J. Bus. Stud.*
- [26] U.S. Equal Employment Opportunity Commission. n.d. *Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures*. EEOC, Washington, DC.
- [27] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* 24, 3, 45-77.
- [28] Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. Design science in information systems research. *MIS Q.* 28, 1, 75-105.
- [29] Touvron, H., et al. 2023. Llama 2: open foundation and fine-tuned chat models. *arXiv:2307.09288*.
- [30] Chung, H. W., et al. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.
- [31] Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 1135-1144. doi:10.1145/2939672.2939778